

Deanship of Graduate Studies

Al-Quds University

**An Evaluation of Ridge Regression in the Presence of
Multicollinearity**

Suliman Sameir Al-Faqeih

M.Sc. Thesis

Jerusalem-Palestine

1432/2011

**An Evaluation of Ridge Regression in the Presence of
Multicollinearity**

By

Suliman Sameir AL-Faqeih

B.Sc.: College of Science and Technology

Al-Quds University/Palestine

Supervisor: Khaled A. Sallah, PhD.

This thesis is submitted in Partial fulfillment of requirements of the degree of Master of Science , Department of Mathematics / Program of Graduate studies.

Al-Quds University

2011

The Program of Graduate Studies /Department of Mathematics
Deanship of Graduate Studies

**An Evaluation of Ridge Regression in the Presence of
Multicollinearity**

By

Student Name: Suliman Al-Faqeih

Registration Number: 20820196

Supervisor: Dr. Khaled A. Sallah

Master thesis submitted and accepted date:

The names and signatures of the examining committee members are as follows

- | | | |
|--------------------------|--------------------|-----------------|
| 1. Dr. Khaled A. Sallah | Head of Committee, | Signatures..... |
| 2. Dr. Tahseen Almograbi | Internal Examiner, | Signatures..... |
| 3. Dr. Fesal Awartani | External Examiner, | Signatures..... |

Al-Quds University

1432/2011

Declaration

I certify that the thesis, submitted for the degree of master, is the result of my own research except where otherwise acknowledged, and that the thesis (or any part of the same) has not been submitted for a higher degree to any other university or institution.

Signed

Suliman Al-Faqeh

Date :

Dedication

To my mother , Sameha

To my father , Sameir

To my brothers , Samer, Mohammed

To my sisters, San'a, Suheir, Suhad, Samya, Sajeda

To my friend

To my colleagues "teachers"

Acknowledgments

This work in this thesis has been carried out at the Department of Mathematics at Al-Quds university. I cannot fully express my gratitude to my supervisor Dr. Khaled A. Sallah. He is an exceptional academic advisor, supporting, extraordinary leader and ready any time to devote his time and efforts to help his students. In the same time he set an example for me by being just, open and honest, kind and gentle person, devoted and caring constant source of wisdom and experience. I really appreciate his patience, encouragement and valuable suggestions.

I am grateful to all the following doctors who taught me during the MA degree: Dr. Yousif Zahalqia, Dr. Ibrahim Qrouz, Dr .Jameel jamal, Dr. Taha Abu kaf, Dr. Abedulhakeem Eidah, Dr. Mohammed Kaleel, Dr. Yousif Bedar, Dr.Tahseen Almougrabi.

Abstract

In regression, the objective is to explain the variation in one or more response variables, by associating this variation with proportional in one or more explanatory variables. If there is no linear relationship between these explanatory variables, they are said to be orthogonal. If the variables is not orthogonal then several of the explanatory variables will vary in rather similar ways. This problem is called multicollinearity, which is a commonly occurring problem in regression analysis. It is the situation in which two or more explanatory variables are highly (but not perfectly) correlated to one another, making it difficult to interpret the strength of the effect of each variable. Handling multicollinearity problem in regression analysis is very important because the least squares estimations assume that the predictors are not correlated. A number of procedures have been developed for finding biased estimators of regression parameters. Some of these procedures are ridge regression (RR), principal component regression (PCR) and partial least squares regression (PLSR).

In this thesis, we consider ridge regression, including the ridge estimator (ordinary and generalized) and their properties. Since the creative work of Hoerl and Kennard ridge regression has proven to be a useful technique to tackle the multicollinearity problem in the linear regression model. Different approaches are investigated with different criteria for estimating the ridge parameter k . In this thesis a comparison between well-known approaches for selecting the ridge parameter k . Under the normality assumptions, the mean squared error (MSE) criterion is used to examine the performance of these estimators when compared with the ordinary least squared estimator (OLS).

The simulation studies and the analysis of real data demonstrate that under certain conditions, at least one of the considered estimators (HKa, KS, HK, FK) have a smaller MSE than the ordinary least squared estimator (OLS), and other approaches.

المخلص

في هذه الرسالة قمنا بدراسة طريقة ما يسمى (Ridge Regression) في ايجاد معاملات المتغيرات المستقلة لحل مشكلة الارتباط الخطي (الازدواج الخطي) بين المتغيرات المستقلة، حيث ان هذه المشكلة تؤثر على طريقة تفسير أي علاقة بين المتغيرات المستقلة والمتغير التابع.

ومن اجل ذلك قمنا بما يلي

قمنا بدراسة ومراجعة خصائص تحليل الانحدار الخطي، من خلال دراسة الانحدار الخطي البسيط والمتعدد وكذلك دراسة ما يسمى (Ordinary least squares) ومعرفة خصائصه ومدى فعاليته في إيجاد معاملات المتغيرات المستقلة ثم قمنا بدراسة مشكلة الارتباط الخطي بين المتغيرات ومراجعة احدث الأبحاث المنشورة عن هذه المشكلة حيث تعرفنا طرق اكتشاف هذه المشكلة ومدى تأثير هذه المشكلة على تفسير العلاقة بين المتغير المستقل والمتغير التابع. قمنا بدراسة احدث الأوراق والأبحاث المنشورة عن (Ridge Regression) ، من خلال هذه الأوراق المنشورة في مجلات عالمية تعرفنا على الكثير من الطرق في إيجاد ما يسمى ridge parameter k .
و قمنا بعمل مقارنة بين ما يسمى

(Mean squared error of OLS وكذلك mean squared error of ridge estimators)

معتمدا على

parameters $\{\hat{k}_{HK\alpha}, \hat{k}_{KS}, \hat{k}_{HK}, \hat{k}_{FK}\}$.

وذلك من خلال توليد مجموعات مختلفة من البيانات تحوي الارتباط الخطي المتعدد بدرجات مختلفة وأظهرنا مدى فاعلية طريقة (Ridge Regression) على (Ordinary least squares) وكذلك التمييز بين قيم k التي اعتمد ridge estimators

وكذلك قمنا بتطبيق بيانات حقيقية على طريقة (Ridge Regression) بالاعتماد على ridge parameters $\{\hat{k}_{HK}, \hat{k}_{KS}, \hat{k}_{HK}, \hat{k}_{FK}\}$.

TABLE OF CONTENTS

List of Tables.....	ix
List of Figures.....	ix
List of Symbols.....	xv
List of Abbreviations.....	xvi

CHAPTER	Page
1 RIDGE REGRESSION IN LITERATURE	1
1.1 Introduction.....	1
1.2 The Problem of Multicollinearity.....	2
1.3 Purpose and Objectives of the Thesis.....	4
1.4 Scope of the Thesis.....	5
2 LINEAR REGRESSION	7
2.1 Simple Linear Regression.....	7
2.1.1 Assumptions on which simple linear regression is based	8
2.1.2 Ordinary Least Square Estimation.....	9
2.1.3 Properties of least squares estimator method.....	10
2.1.4 Analysis of Variance.....	10
2.1.5 Tests of hypotheses.....	11
2.1.5.1 Test of β_1	12
2.1.5.2 Explanation for testing β_1	12
2.1.5.3 test of β_0	13
2.1.6 Confidence Interval.....	13
2.1.7 Coefficient of determination	14
2.3 Multiple linear regression.....	14
2.3.1 Estimation of regression coefficients.....	15
2.3.2 Proprieties of the ordinary least estimator.....	16
2.3.3 Distribution of \mathbf{b}	17
2.3.4 Tests for β_k	18

2.3.5	Confidence interval for β_k	18
2.4	Correlation transformation.....	19
3	Multicollinearity.....	21
3.1	Explanation of multicollinearity.....	22
3.2	Effects of multicollinearity	23
3.3	Source of multicollinearity.....	24
3.4	Multicollinearity diagnostics.....	25
3.4.1	Informal diagnostics.....	25
3.4.2	formal diagnostics.....	27
3.4.2.1	Variance inflation factor (VIF).....	27
3.4.2.2	Tolerance	29
3.4.3.2	Eigenvalues, Conditions numbers	29
3.5	Remedies of multicollinearity.....	30
3.5.1	Model respecification.....	30
3.5.2	Use additional or new data.....	31
3.5.3	Principal component regression.....	32
4	Ridge Regression.....	35
4.1	Ridge regression estimators	35
4.2	Selection of variables in ridge regression by ridge trace.....	40
4.3	General ridge regression.....	40
4.3.1	General ridge regression (I).....	41
4.3.2	Superiority of the GRR $\hat{\alpha}^{GR}$ over OLS $\hat{\alpha}$	43
4.3.3	General ridge regression(II).....	45
4.3.4	Superiority of $\hat{\alpha}^*$ over the ordinary least square $\hat{\beta}$	46
4.4	Ridge parameter k	48
5	Applications.....	52
5.1	Simulated data.....	52
5.1.1	Generating Simulated Data Sets	52
5.1.2	Performance of Ridge Regression to simulated data.....	56
5.1.3	Simulation results.....	58
5.2	Real data.....	63

5.2.1	Data base.....	63
5.2.2	Data analysis.....	64
5.2.3	Performance of Ridge Regression to real data.....	69
5.3	Summery and conclusions	71
References		73
Appendix		78
	Appendix A.....	78
	Appendix B.....	80
	Appendix C.....	83

List of Tables

Table (4.1) Ridge parameters which we made a comparison between them.

Table (5.1) Factors and levels for the simulated data sets .

Table (5.2) Group one of simulated data.

Table (5.3) Group two of simulated data.

Table (5.4) Group three of simulated data.

Table (5.5) The value of correlation for $p = 10, n = 15$.

Table (5.6) The value of correlation for $p = 10, n = 30$.

Table (5.7) The value of correlation for $p = 10, n = 50$.

Table (5.8) The value of correlation for $p = 10, n = 80$.

Table (5.9) The value of correlation for $p = 5, n = 30$.

Table (5.10) The The value of correlation for $p = 2, n = 30$.

Table (5.11) Value of correlation for $p = 2, n = 15$.

Table (5.12) The value of correlation for $p = 2, n = 50$.

Table (5.13) Estimated MSE for group one of simulated data.

Table (5.14) Estimated MSE for group two of simulated data.

Table (5.15) Estimated MSE for group three of simulated data.

Table (5.16) The selected variables of the vehicle characteristics.

Table (5.17) Correlation Coefficients between deferent variables.

Table (5.18) Checking the Model Fit (ANOVA).

Table (5.19) Model Summary of real data.

Table (5.20) Model Coefficients of real data.

Table (5.21) Collinearity Diagnostics of real data.

Table (5.22) estimated MSEs of real data.

Table (5.23) estimated ridge coefficient for real data.

List of Figures

Figure (3.1) The choice of VIF value against the R-square value.

Figure (3.2) Steps in PCR algorithm.

Figure (4.1) The sampling distribution of biased and unbiased estimator.

Figure (5.1) Flowchart summarizing performance of RR.

Figure (5.2) the steps in ridge regression algorithm.

Figure (5.3) Histogram with normal probability plot of the residuals.

Figure (5.4) Normal P-P Plot of Regression Standardized Residual.

Figure (5.5) Steps in ridge regression algorithm for real data.

LIST OF SYMBOLS

Response (dependent) variable	Y
Predictor (independent) variable	X
Parameter (regression coefficient), known constant	B
Variance	σ^2
Matrix of predictors	\mathbf{X}
Vector of parameters	$\boldsymbol{\beta}$
Vector of error matrix term	$\boldsymbol{\varepsilon}$
Identity matrix	\mathbf{I}
Estimate regression coefficient	$\hat{\beta}$
Vector of estimate regression coefficient	$\hat{\boldsymbol{\beta}}$
Fitted value	\hat{y}
Residual term	E
Vector of residual term	\mathbf{e}
Number of observations	N
Number of regressors	P
Shrinkage parameter	K
Vector of Shrinkage parameter	\mathbf{k}
Matrix of eigenvectors	\mathbf{V}
The diagonal matrix of eigenvalues	$\boldsymbol{\Lambda}$
Coefficient of Determination	R^2
Components for Principal Component Regression	\mathbf{Z}
Eigenvalue	Λ
Correlation matrix	r_{XX}
Correlation between X and Y	r_{XY}

List of Abbreviations

ABBREVIATIONS	MEANING
CN	Condition Number
Cov	Covariance
LS	Least Squares
Max	Maximum
MSE	Mean Square Error
OLS	Ordinary Least Squares
SSTO	Total sums of squares
SSE	Error sums of squares
SSR	Regression sum of square
PCR	Principal Component Regression
RR	Ridge Regression
VIF	Variance Inflation Factors
GR	General Ridge Regression

Chapter one

INTRODUCTION

1.1 Introduction

Researchers are often interested in the relationships between one variable and several other variables. Often in applied statistics, after the data had been collected the purpose of analysis is to construct a statistical model. Regression analysis consists of techniques for modeling the relationships between a dependent variable (known as response variable) and one or more independent variables (known as explanatory variables or predictors). In regression the dependent variable is modeled as a function of independent variables, corresponding regression parameter, and a random error term which represents the variation in the dependent variable unexplained by the function of the independent variables in symbol we denote the response variable by Y and the set of predictor variables by X_1, X_2, \dots, X_p where p denotes the number of predictor variables. The relation between Y and the set of independent variable X_1, X_2, \dots, X_p can be approximated by the regression model

$$Y = f(X_1, X_2, \dots, X_p) + \varepsilon$$

Where ε is assumed to be a random error (noise weight) representing the discrepancy in the approximation. It accounts for the failure of the model to fit the data exactly. The function $f(X_1, X_2, \dots, X_p)$ describes the relationship between Y and X_1, X_2, \dots, X_p . f can be linear or nonlinear function. The term linear (nonlinear) doesn't describe the relationship between Y and X_1, X_2, \dots, X_p . It is related to the fact that the regression parameters enter the equation

linearly (nonlinearly). Linear regression requires that the model is linear in regression parameters. Regression analysis is the method to discover the relationship between one or more response variables and the predictors. There are three types of regression. The first is simple linear regression. The simple linear regression is for models the linear relationship between two variables one of them is the dependent variable and the other is the independent variable. The second type in regression is the multiple linear regression which is linear regression model with one dependent variable and more than one independent variables. The third type of regression is nonlinear regression, which assumes that the relationship between the dependent variable and the independent variable is not linear in regression parameters. Nonlinear regression model is more complicated than linear regression model in term of estimation the model parameters, model selection, model diagnosis, variable selection, outlier observation. When we deal only with one response variable, regression analysis is called univariate regression and in case we have two or more response variables regression is called multivariate regression.

1.2 The Problem of Multicollinearity

The problem of multicollinearity has remained the center of attraction in the literature of linear regression analysis for a long time. It arises when the explanatory variables in a linear regression model are highly correlated, and thus one or more columns of the (design matrix) form a near linear combination with other columns. This problem can cause the value of the least squares estimated regression coefficients to be conditional upon the correlated predictor variables in the model. As defined by Bowerman and O'Connell (1990) the presence of multicollinearity in the data is a numerical issue as well as a statistical issue. It is a statistical issue because it inflates (being large) the variance of ordinary least

square estimator and a numerical issue in the sense that the small errors in input may cause large errors in the output.

According to Neter et al. (1996), there are two types of multicollinearity: perfect multicollinearity (or extreme multicollinearity) and high multicollinearity (or near extreme multicollinearity).

Perfect multicollinearity means that at least two of the independent variables in a regression equation are perfectly related by a linear function. When perfect multicollinearity is present, there is no perfect solution. Perfect multicollinearity occurs when:

1. Independent variables are linear functions of each other, for example; age and year of birth.
2. Dummy variables are created for all values of a categorical variable.
3. There are fewer observations than variables.

High multicollinearity means that there are strong (but not perfect) linear relationship among the independent variables. If the regression model has only two independent variables, high multicollinearity occurs if the two variables have a correlation that is close to 1 or -1 . Therefore, the closer it gets to 1 or -1 , the greater is the association between the independent variables. When there is high but imperfect multicollinearity, a solution is still possible but as the independent variables increase in correlation with each other, the standard errors of the regression coefficients will become inflated.

1.3 Purpose and Objectives of the Thesis

A recent alternative to least squares regression is ridge regression (Darlington, 1978; Dempster, Schatzoff & Wermuth, 1977; Hoerl & Kennard, 1970; Prince, 1977). Ridge regression was developed expressly for the purpose of circumventing the weakness of least squares regression with regard to highly overlapping predictors, and as such, applying ridge regression would appear to be very appropriate.

It is important to realize that the resulting ridge regression equation is a biased estimate and not reflective of population parameters. As such, ridge regression is of little use in theoretical modeling (Darlington, 1978). The main advantage of ridge regression is in prediction, and as this is the specific purpose of using regression equations in selection, ridge regression would appear to be a particularly useful tool when multicollinearity is presented.

The purpose of this thesis is to examine the improvement in prediction of the ridge regression procedure over the least squares regression procedure when applied to multicollinearity data. Improvement of prediction was defined in terms of the value of ridge parameters $\hat{k}_{HK\alpha}$, \hat{k}_{KS} , \hat{k}_{HK} and, \hat{k}_{KF} of the multiple correlation coefficients obtained when the regression equations were applied on different samples. Further, as it has been demonstrated that the best choice of k gives better maximum prediction. It was expected that the ridge regression procedure would result in less MSE than the least squares procedure.

The objectives of this thesis can be summarized as:

1. To study and investigate the univariate and multivariate linear regression and their properties.
2. To present the ridge estimator (ordinary and generalized) and its properties for handling multicollinearity problem.
3. To review the relevant literature on published work done recently concerning the problems of multicollinearity.
4. To compare the mean squared error of OLS and the mean squared error of ridge estimators depend on the ridge parameters $\{\hat{k}_{HK\alpha}, \hat{k}_{KS}, \hat{k}_{HK}, \hat{k}_{FK}\}$

1.4 Scope of the Thesis

This thesis consists of five chapters. The first chapter is an introductory chapter in which the definition of regression analysis, multicollinearity and the importance of the ridge regression are mentioned.

In the second chapter, some background information about the linear regression is presented. In addition, some common estimate tasks and techniques are explained, and the main steps of the Ordinary Least Square Estimation (OLSE) process are mentioned. At the end of this chapter, some popular approaches of statistical inferences of the (OLSE) are presented.

The main part of this thesis contains three chapters, which cover research related to this study. In the third chapter a comprehensive literature review on multicollinearity problem is introduced. Besides the effect, the source, the diagnostics, and the remedies of

multicollinearity are presented. A detailed explanation of ridge regression, selection of variables, general ridge regression and ridge parameters will be presented in chapter four. A comparison study through a simulation study and real data analysis followed by conclusion of this study and the possible works that can be done in the future are presented in the final chapter.

Chapter two

Linear Regression

This chapter expands on the analysis of simple linear regression models and discusses the analysis of multiple linear regression models in matrix form.

Numerous procedures have been developed for parameter estimation and inference in linear regression. These methods differ in computational simplicity of algorithms, presence of a closed-form solution, robustness with respect to heavy-tailed distributions, and theoretical assumptions needed to validate desirable statistical properties such as consistency and asymptotic efficiency. One of the most common estimation techniques for linear regression is the Ordinary Least Squares (OLS). Discussion and properties of the (OLS) also are presented in this chapter.

2.1 Simple Linear Regression

The general form of simple linear regression consists of the mean function and the variance function

$$\begin{aligned} E(Y|X = x) &= \beta_0 + \beta_1 x \\ \text{Var}(Y|X = x) &= \sigma^2 \end{aligned} \tag{2.1}$$

The parameters in model (2.1) are the intercept β_0 which is the value of $E(Y|X = x)$ when x is equals zero and β_1 is the slope which is the rate of change in $E(Y|X = x)$ for a unit change in X . As this parameter changes we get different straight lines. In most applications parameters are unknown and must be estimated. σ^2 is constant that is usually unknown. The observed value of the response variable y_i will be typically not equal to the expected

value $E(Y|X = x)$ since $\sigma^2 > 0$. The difference between the observed value and the expected value is called statistical error or ε_i , that is

$$\varepsilon_i = y_i - E(Y|X = x_i), \quad i = 1, 2, \dots, n$$

The errors ε_i depend on the unknown parameter in the mean function and so are not observable quantities. They are random variables and correspond to vertical distance between y_i and $E(Y|X = x)$.

2.1.1 Assumptions on which simple linear regression is based

Quantitative models always rest on assumptions about the way the world works, and regression models are no exception. There are four principal assumptions which justify the use of linear regression models for purposes of prediction:

1. The mean value of the dependent variable Y increases or decreases linearly as the value of the independent variable X increases or decreases. To put it simply there is a linear relationship between X and Y .
2. For given value of independent variable X the corresponding values of the dependent variable Y are distributed normally. The mean value of this distribution falls on the regression line.
3. The standard deviation of the values of the dependent variable Y at any given value of independent variable X is the same for all values of X .
4. The errors are uncorrelated (i.e. Independent) and are distributed normally as

$$\varepsilon_i \sim N(0, \sigma^2)$$

2.1.2 Ordinary Least Square Estimation

There are many methods for estimating the parameters in the model (2.1). Here we will discuss the ordinary least squares method (OLS), in which parameter estimates are chosen to minimize the residual sums of squares. Estimates of parameters are computable functions of data and therefore statistics, we estimate β_1 by $\hat{\beta}_1$, and β_0 by $\hat{\beta}_0$ and thus the model (2.1) is estimated as

$$\hat{E}(Y|X = x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

The best linear model minimizes the sum squared error (RSS):

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Estimation of a simple linear regression relationship involves finding estimated or predicted values of the intercept and slope of the linear regression line.

The estimated regression line is:

$$\hat{y}_i = \hat{E}(Y|X = x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, 2, \dots, n$$

The least squares estimates can be derived in many ways. And they are given by the expressions:

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \quad \text{And} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

$$\text{where } S_{XY} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{n-1}} \quad \text{And} \quad S_{XX} = \frac{\sum (x_i - \bar{x})^2}{\sqrt{n-1}}$$

2.1.3 Properties of least square method

With the assumptions on the error that assumed to be independent random quantities and Normally distributed with mean zero and variance σ^2 the fitted line pass through the point (\bar{x}, \bar{y}) . The quantities $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimates of β_0 and β_1 respectively. Their variances are:

$$Var(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right] \quad \text{and} \quad Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

And they are correlated, with covariance

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{\sum(x_i - \bar{x})^2}$$

Furthermore, the sampling distribution of the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are Normal with means β_0 and β_1 and variances that given above respectively.

The variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ depend on the unknown parameter σ^2 from the data. An unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n - 2}$$

2.1.4 Analysis of Variance

Analysis of variance provides a convenient method of comparing the fit of two or more mean functions for the same set of data. The total sum of squared deviations in Y can be decomposed into the sum of two quantities the first, SSR, measures the quality of X as a

predictor of Y , and the second, SSE, measures the error in this prediction. This is explained in the following points:

- Our observed variable Y will always have some variability associated with it. We break this into the variability related to our predictor and the variability unrelated to our predictor.
- SSTO=Total sums of squares= $\sum(y_i - \bar{y})^2$ is the total variability in Y . It has $n - 1$ degrees of freedom associated with it.
- SSR=regression sum of square = $\sum(\hat{y}_i - \bar{y})^2$ is the variability of Y accounted for our regression model. Since we are using one predictor X it has one degree of freedom associated with it.
- SSE=Error sums of squares = $\sum(y_i - \hat{y}_i)^2$ is the variability in Y that is not accounted by for our regression model. It has $n - 2$ degree of freedom associated with it.
- The fundamental equality is given by

$$SSTO = SSE + SSR$$

2.1.5 Tests of hypotheses

To test whether there is a linear relationship between two variables we can perform a hypothesis test on the slope parameter in the corresponding simple linear regression model.

The general form for test statistic is

$$\frac{\text{point estimate} - E(\text{point estimate})}{\text{standard deviation of point estimate}}$$

2.1.5.1 Test of β_1

In order to test β_1 , we consider the following point :

- Point estimate (β_1) = least squares estimate = $\hat{\beta}_1$
- E(point estimate) = β_1
- Standard deviation = $s.e(\hat{\beta}_1) = \sqrt{Var(\hat{\beta}_1)} = \sqrt{\frac{\sigma^2}{\sum(x_i - \bar{x})^2}}$
- The resulting test statistic $t_1 = \frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)}$, follows a t distribution with $n - 2$ degree of freedom.

2.1.5.2 Explanation for testing β_1

An appropriate test statistic for testing the null hypothesis $H_0: \beta_1 = 0$ against the alternative $H_1: \beta_1 \neq 0$ in the t -test,

$$t_1 = \frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)}$$

The statistic t_1 is distributed as Student's t -distribution with $(n - 2)$ degrees of freedom.

The test is carried out by comparing this observed value with the appropriate critical value obtained from the t -table which is $t_{(n-2, \alpha/2)}$ where α is specified significance level.

Accordingly H_0 is rejected at the significance level α if

$$|t_1| \geq t_{(n-2, \alpha/2)} \quad (2.2)$$

Where $|t_1|$ denotes the absolute value of t_1 . A criterion equivalent to that in (2.2) is to compare the p -value for the t_1 -test with α , and we reject H_0 if

$$p(|t_1|) \leq \alpha$$

2.1.5.3 Test of β_0

In order to test β_0 we consider the following point:

- Point estimate = least squares estimate = $\hat{\beta}_0$
- $E(\text{point estimate}) = \beta_0$
- Standard deviation = $s.e(\hat{\beta}_0) = \sqrt{\text{Var}(\hat{\beta}_0)} = \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right]}$
- The resulting test statistic $t_1 = \frac{\hat{\beta}_0 - \beta_0}{s.e(\hat{\beta}_0)}$, follows a t distribution with $n - 2$ degrees of freedom.

2.1.6 Confidence Interval

Point estimate tell us about the central tendency of a distribution while confidence intervals tell us about both the central tendency and the spread of the distribution.

The general form of the confidence interval is

Point estimate \mp (critical value)(standard error of point estimate)

The general $(1 - \alpha)$ confidence interval for β_1 is given by:

$$\hat{\beta}_1 \mp \left(t_{(n-2, 1-\frac{\alpha}{2})} \right) s.e\{\hat{\beta}_1\}$$

We can perform significance test using confidence intervals. If your interval contains the value of the parameter under the null hypothesis you fail to reject the null. If the value under the null hypothesis falls outside of your confidence interval then you reject the null.

2.1.7 Coefficient of determination

The Coefficient of determination r^2 , indicate the percentage of variation in Y that is explained by all predictor in the equation. The Coefficient of determination can be calculated as

$$r^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

Properties of r^2 summarized as follows:

- $0 \leq r^2 \leq 1$
- Higher of r^2 leads to more useful model.
- Unaffected if the units of the measurement are changed.
- The Coefficient of determination r^2 is a measure of how well the least square model perform as a predictor of Y .
- The Coefficient of determination r^2 measures the relative size of $SSTO$ and SSE .

2.3 Multiple linear regression

A multiple linear regression model can be expressed as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \quad (2.3)$$

Model (2.3) can be written in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2.4)$$

Where \mathbf{y} is $n \times 1$ vector of the variable to be explained, \mathbf{X} is an $n \times p$ matrix of explanatory variables where n is the number of observations and p is the number of the

explanatory variable. $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of disturbances distributed as $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$. The $p \times 1$ parameter vector $\boldsymbol{\beta}$ is assumed unknown and to be estimated by the data \mathbf{y} and \mathbf{X}

Where

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p-1} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{np-1} \end{pmatrix},$$

$$\mathbf{y} = (Y_1, \dots, Y_n)^t,$$

$$\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^t$$

And

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{p-1})^t$$

2.3.1 Estimation of regression coefficients

The least squares criterion is generalized for general linear regression model as

$$SSE = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - \dots - \beta_p X_{ip-1})^2$$

The least squares estimators are those values of $\beta_0, \beta_1, \dots, \beta_{p-1}$ that minimize the sum square error. Denote the vector of least squares estimated by

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}$$

In matrix form the ordinary least squares written as

$$\begin{aligned}
SSE &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{y}\mathbf{y}^t - \mathbf{y}^t\mathbf{X}\mathbf{b} - \mathbf{b}^t\mathbf{X}^t\mathbf{y} + \mathbf{b}^t\mathbf{X}^t\mathbf{X}\mathbf{b} \\
&= \mathbf{y}\mathbf{y}^t - 2\mathbf{y}^t\mathbf{X}\mathbf{b} + \mathbf{b}^t\mathbf{X}^t\mathbf{X}\mathbf{b}
\end{aligned}$$

Taking the derivative with respect to \mathbf{b} gives,

$$\frac{\partial SSE}{\partial \mathbf{b}} = 0 - 2\mathbf{y}^t\mathbf{X} + 2\mathbf{X}^t\mathbf{X}\mathbf{b}$$

Setting this equal to zero implies

$$\mathbf{X}^t\mathbf{X}\mathbf{b} = \mathbf{X}^t\mathbf{y}$$

This is called the normal equation

Assuming that $\mathbf{X}^t\mathbf{X}$ not ill-conditioned matrix thus we have the unique linear solution

$$\mathbf{b} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

Thus the predicted model is given by

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

2.3.2 Proprieties of the ordinary least estimator

- The Gauss-Markov theorem tells us that the OLS estimator is the best unbiased linear Estimator (BUE).
- Unbiased means that the expected value of \mathbf{b} , $E(\mathbf{b}) = \boldsymbol{\beta}$.
- The estimator is linear function of the dependent variable observation once we have fixed the model matrix \mathbf{X} .
- The least squares estimator is under the assumption, the best such estimator because it is the most efficient (minimum variance).

- The OLS is attractive because it is Maximum likelihood estimator.

2.3.3 Distribution of \mathbf{b}

Since $\mathbf{b} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$, the distribution of \mathbf{b} is based on the distribution of \mathbf{y}

Since $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ and by multivariate theorem we have

$$E(\mathbf{b}) = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^tE\mathbf{y} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

$$\sigma^2\{\mathbf{b}\} = cov\{\mathbf{b}\} = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$$

Thus

$$\mathbf{b} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}) \quad (2.5)$$

We can generalize what we mentioned in section (2.1.4) as following:

- $SSTO = \sum(Y_i - \bar{Y})^2$, and it has $n - 1$ degrees of freedom associated with it.
- $SSR = \sum(\hat{Y}_i - \bar{Y})^2$, and it has $p - 1$ degree of freedom associated with it.
- $SSE = \sum(Y_i - \hat{Y}_i)^2$, and it has $n - p$ degree of freedom associated with it.

And the mean squares:

$$MSR = \frac{SSR}{p - 1} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{p - 1}$$

$$MSE = \frac{SSE}{n - p} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n - p}$$

$$MST = \frac{SSTO}{n - 1} = \frac{\sum(Y_i - \bar{Y})^2}{n - 1}$$

2.3.4 Tests for β_k

The test for significance of regression is to test to determine whether a linear relationship exists between the response variable and the independent variable. The appropriate hypotheses are

$$H_o : \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$$

$$H_1 : \beta_k \neq 0 \text{ for at least one } k = 1, \dots, p - 1$$

The test statistic for $H_o : \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$ is

$$F = \frac{MSR}{MSE}$$

Under H_o , $F \sim F_{p-1, n-p}$

We Reject H_o if F is greater than critical value $F_{p-1, n-p}$. If H_o is rejected, we conclude that at least one of the regression coefficients is non zero hence at least ones of the X variable is useful in predicting Y. If H_o is not rejected, then we cannot conclude that any of the X variables is useful in predicting Y.

2.3.5 Confidence interval for β_k

We can construct confidence intervals for a particular coefficient β_k . The $1 - \alpha$ confidence interval is given by

$$b_k \pm t_{(n-p, 1-\frac{\alpha}{2})} s\{b_k\}$$

From (2.5) we get that

$$s^2\{\mathbf{b}\}_{p \times p} = MSE \times (\mathbf{X}^t \mathbf{X})^{-1}$$

Thus

$$s\{b_k\} = \sqrt{[s^2\{\mathbf{b}\}]_{k \times k}}$$

2.4 Correlation transformation

The Correlation transformation is a simple function of the standardized variables, the Correlation transformation of the dependent and independent variables is given by

$$\tilde{X}_{ik} = \frac{1}{\sqrt{n-1}} \left(\frac{X_{ik} - \bar{X}_k}{\hat{\sigma}_X} \right), k = 1, 2, \dots, p-1$$

$$\tilde{Y}_i = \frac{1}{\sqrt{n-1}} \left(\frac{Y_i - \bar{Y}}{\hat{\sigma}_Y} \right), i = 1, 2, \dots, n$$

The regression model with transformed variables as defined by the correlation transformation is called a standardized regression model and defined as follows :

$$\tilde{Y}_i = \tilde{\beta}_1 \tilde{X}_{i1} + \tilde{\beta}_2 \tilde{X}_{i2} + \dots + \tilde{\beta}_{p-1} \tilde{X}_{i,p-1} + \tilde{\varepsilon}_i$$

The \mathbf{X} matrix for the transformed variables (without the intercept term) is

$$\mathbf{X} = \begin{pmatrix} \tilde{X}_{11} & \dots & \tilde{X}_{1,p-1} \\ \vdots & \dots & \vdots \\ \tilde{X}_{n1} & \dots & \tilde{X}_{n,p-1} \end{pmatrix}$$

Then $\mathbf{X}^t \mathbf{X} = \mathbf{r}_{XX}$, where \mathbf{r}_{XX} is the correlation matrix of the X variables which contains the element of coefficients of simple correlation between all pairs of X variables. That is,

$$\mathbf{r}_{XX} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1,p-1} \\ \vdots & 1 & \ddots & \vdots \\ r_{1,p-1} & r_{2,p-1} & \dots & 1 \end{bmatrix}$$

Similar to the algebraic definition of $\mathbf{X}^t \mathbf{X}$ matrix

$$\mathbf{X}^t\mathbf{Y} = \mathbf{r}_{XY}$$

Where

$$\mathbf{r}_{XY} = \begin{pmatrix} r_{Y1} \\ r_{Y2} \\ \cdot \\ \cdot \\ r_{Yp-1} \end{pmatrix}$$

The normal equation for the standardized multiple regression is given by

$$\tilde{\mathbf{b}} = (\mathbf{r}_{XX})^{-1}\mathbf{r}_{XY}$$

Where

$$\tilde{\mathbf{b}} = \begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \\ \cdot \\ \cdot \\ \tilde{\beta}_{p-1} \end{pmatrix}$$

The parameters $\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_{p-1}$ in the standardized regression model and the original parameters $\beta_0, \beta_2, \dots, \beta_{p-1}$ in the ordinary multiple regression model are related as follows :

$$\beta_i = \left(\frac{\hat{\sigma}_Y}{\hat{\sigma}_X} \right) \tilde{\beta}_i, i = 1, 2, \dots, p - 1$$

$$\beta_0 = \bar{Y} - \beta_1\bar{X}_1 - \beta_2\bar{X}_2 - \dots - \beta_{p-1}\bar{X}_{p-1}$$

Chapter three

Multicollinearity

If there is no linear relation between the predictors, then they are said to be orthogonal. When the predictors are orthogonal, prediction of dependent variable and estimation of the parameters coefficient and selection an appropriate predictor in the model can be made relatively easily. Unfortunately, in most application on linear regression the independent variables are not orthogonal that is there are approximate linear relationships between two or more independent variables in a multiple regression model. When there are near linear dependencies between predictors multicollinearity exists. The condition of severe nonorthogonality is also referred to as the problem of collinear data, or multicollinearity.

Elimination of multicollinearity is not possible completely but the degree of multicollinearity can be decreased. In this chapter our study will be focused on explaining multicollinearity problem. And discusses several methods for detection this problem such as variance inflation factor VIF, correlation matrix, condition number, and tolerance. Some of the popular methods for decreasing the degree of multicollinearity such as principal component regression, adding additional data or new data, model respecification will be discussed in this chapter. The most popular method for handling multicollinearity problem is ridge regression, this method will be discussed in details in chapter four.

3.1 Explanation of Multicollinearity

In most applications perfect multicollinearity is unlikely but near multicollinearity is more likely to analyst. Let j^{th} column of the matrix $\mathbf{X} = [\mathbf{X}_1 \mathbf{X}_2 \dots \mathbf{X}_j \dots \mathbf{X}_p]$ denoted by \mathbf{X}_j , multicollinearity can be defined as the linear dependence of the column of \mathbf{X} . The vectors are linearly dependent if there is a constant c_1, c_2, \dots, c_p not all equal zero such that

$$\sum_{j=1}^p c_j \mathbf{X}_j = 0 \quad (3.1)$$

If (3.1) holds for a subset of columns of \mathbf{X} then the rank of $\mathbf{X}^t \mathbf{X}$ is less than p and $(\mathbf{X}^t \mathbf{X})^{-1}$ doesn't exist, and if (3.1) holds approximately for some subsets of \mathbf{X} , then there will be a near linear dependency in $\mathbf{X}^t \mathbf{X}$ and the problem of multicollinearity exists. It is to be noted that the multicollinearity is a form of ill-conditioning in the $\mathbf{X}^t \mathbf{X}$ matrix. Furthermore, the problem is one of the degrees, that is, every data set will suffer from multicollinearity to some extent unless the columns of \mathbf{X} are orthogonal.

The presence of multicollinearity can make the usual least squares analysis of the regression model dramatically inadequate. In some cases, multiple regression results may seem paradoxical. Even though the overall P value is very low, all of the individual P values are high. This means that the model fits the data well, even though none of the X variables has a statistically significant impact in predicting Y . How is this possible? When two X variables are highly correlated, they both convey essentially the same information. In this case, neither may contribute significantly to the model after the other one is included. But together they contribute a lot. If both variables are removed from the model, the fit would be much worse. So the overall model fits the data well, but neither X variable

makes a significant contribution when its added to the model. When this happens, the X variables are collinear and the results show multicollinearity.

3.2 Effects of multicollinearity

If the goal is simply to predict Y from a set of X variables, then multicollinearity is not a problem. The predictions will be still accurate, and the overall R^2 (or adjusted R^2) quantifies how well the model predicts the Y values. If the goal is to understand how the various X variables impact Y , then multicollinearity is a big problem.

The effects of multicollinearity can be listed as follows:

1. For variables that are highly related to one another but not perfectly related the ordinary least squares estimators have large variances and covariances making precise estimation difficult.
2. Confidence intervals tend to be much wider, the confidence interval may include zero, leading to the acceptance of the null hypothesis more readily which means one can't even be confident whether an increase in the X value is associated with an increase, or a decrease, in Y . Because the confidence intervals are so wide, excluding a subject (or adding a new one) can change the coefficients dramatically and may even change their signs.
3. Although the t ratio of one or more of the coefficients is more likely to be insignificant with multicollinearity, the R^2 value for the model can still be relatively high.
4. The ordinary least squares estimators and their standard errors can be sensitive to small changes in the data. In other words, the results will not be robust.

5. The individual P values can be misleading (a P value can be high, even though the variable is important).
6. Roundoff error in normal equation calculation

The results from normal equations calculations can be sensitive to rounding of data in intermediate stage of calculation. The roundoff errors tend to enter least squares calculations when the inverse of $\mathbf{X}^t\mathbf{X}$ is taken. It may be serious when $\mathbf{X}^t\mathbf{X}$ has a determinant that is close to 0, in which case $(\mathbf{X}^t\mathbf{X})^{-1}$ almost does not exist. This results in inaccurate values of least squares estimated regression coefficients \mathbf{b} . Roundoff error may also exist when the element of $\mathbf{X}^t\mathbf{X}$ differ substantially in terms of magnitude, that is when the data in \mathbf{X} variables cover large range.

Correlation transformation helps with controlling roundoff error because it makes all entries in the $\mathbf{X}^t\mathbf{X}$ matrix for the transformed variable to fall between -1 and $+1$ inclusive. Hence, the calculation of the inverse matrix becomes much less subjected to roundoff error due to dissimilar orders of magnitudes than with the original variables.

3.3 Source of multicollinearity

There are four primary sources of multicollinearity:

1. The data collection method employed. This method can lead to multicollinearity when the analyst samples only a subspace of the region of the regressors defined in equation (2.4).
2. Constraints on the model or the population. Constraints of the model or in the population being sampled can cause multicollinearity. For example of constraints

physical constraints such as the unit of the regressors. And other constraints that the researchers added to the model.

3. Model specification. Multicollinearity may be induced by the choice of model. We know that adding a polynomial term to a regression model causes ill conditioning of the $\mathbf{X}^t\mathbf{X}$ matrix.
4. An over defined model. An over defined model has more regressor variables than number of observations. These models are sometimes uncouneted in medical and behavioral research, where there may be only small number of subjects available, and information is collected for a large number of regressors on each subject.

3.4 Multicollinearity diagnostics

Multicollinearity is a matter of degree, not a matter of presence or absence. The higher degree of multicollinearity, the greater the likelihood of the disturbing consequences of multicollinearity.

There are several techniques that have been proposed for detecting multicollinearity:

3.4.1 Informal Diagnostics

A variety of informal diagnostics can be used to detect multicollinearity problems. These informal diagnostics can be listed as follows:

1. A very simple measure of multicollinearity is inspection of the off-diagonal elements r_{ij} in \mathbf{r}_{XX} . If regressors X_i and X_j are nearly dependent, then $|r_{ij}|$ will be near unity.

2. The determinant of \mathbf{r}_{XX} can be used as an index of multicollinearity, the possible range of values of the determinant is $0 \leq |\mathbf{r}_{XX}| \leq 1$. If $|\mathbf{r}_{XX}| = 1$, the regressors are orthogonal, while if $|\mathbf{r}_{XX}| = 0$, there is an exact linear dependence among regressors. The degree of multicollinearity becomes more severe as $|\mathbf{r}_{XX}|$ approaches zero. While this measure of multicollinearity is easy to apply, it doesn't provide any information on the source of the multicollinearity.
3. The F statistics for significance of regression and individual t statistics can sometimes indicate the presence of multicollinearity. Specifically, if the overall F statistic is significant but the individual t statistics are all non significant, multicollinearity is present. Unfortunately, many data sets that have significant multicollinearity will not exhibit this behavior, and so the usefulness of this measure is questionable.
4. The sign and magnitude of the regression coefficients will sometimes provide an indication that multicollinearity is present. In particular if adding or removing a regressor produces large changes in the estimates of the regression coefficients, multicollinearity is indicated. If the deletion of one or more data points results in large changes in the regression coefficients, there may be multicollinearity present. Finally if the signs or magnitude of the regression coefficients in the regression model are contrary to the prior expectation, we should be alert to possible multicollinearity.
5. The wide confidence intervals for regression coefficients of important predictor variables is also another sign that multicollinearity is present in the regression analysis.

6. Multicollinearity can also cause large changes in the least squares estimated regression coefficients when a predictor variable is added or deleted or when observation is altered or deleted.

The informal methods just described have important limitations. They don't provide quantitative measurements of the impact of multicollinearity and they may not identify the nature of the multicollinearity. Also sometime the observed behavior may occur without multicollinearity being present.

3.4.2 Formal Diagnostics

The development of formal methods for detecting multicollinearity problem is to determine how serious the problem affects the analysis and to know the details of which variables are correlated and need to be omitted or deleted.

3.4.2.1 Variance inflation factor (VIF)

Variance Inflation Factors is the measure of the speed with which variances and covariances increase and it is most commonly used method for detecting multicollinearity problem. Variance inflation factors is a measure of multicollinearity in a regression design matrix (that is, independent variables) in a scaled version of the multiple correlation coefficient between an independent variable, and the rest of the independent variable. The measure shows the number of times the variances of the corresponding parameter estimate is increased due to multicollinearity as compared to as what it would be if there were no multicollinearity. Therefore, this diagnostic is designed to indicate the strength of the linear dependencies and how much the variances of each regression coefficient is inflated above ideal (Myers, 1986).

The diagonal elements of the inverse of the \mathbf{r}_{XX} matrix are very useful for detecting multicollinearity. The j^{th} diagonal element of $\mathbf{C} = (\mathbf{r}_{XX})^{-1}$ matrix can be written as $C_{jj} = (1 - R_j^2)^{-1}$, where R_j^2 is the coefficient of determination obtained when X_j is regressed on the remaining $p - 1$ regressors. If X_j is nearly orthogonal to the remaining $p - 1$ regressors, R_j^2 is small and C_{jj} is close to unity, while if X_j is nearly linearly dependent on some subset of the remaining regressors, R_j^2 is near unity and C_{jj} is large. Since the variance of the j^{th} regression coefficient is $C_{jj}\sigma^2$, we can view C_{jj} as the factor by which the variance of the $\hat{\beta}_j$ is increased due to near linear dependences among the regressors. We call this variance inflation factor or VIF and denoted for each $j = 1, \dots, p$,

$$VIF_j = (1 - R_j^2)^{-1}$$

There is no formal cutoff value to use with the VIF for determining the presence of multicollinearity but, Neter et al. (1996) recommended looking at the largest VIF value. A value greater than 10 is often used as an indication of potential multicollinearity problem. The cutoff value of VIF that should be used to determine whether collinearity is a problem is shown as follows

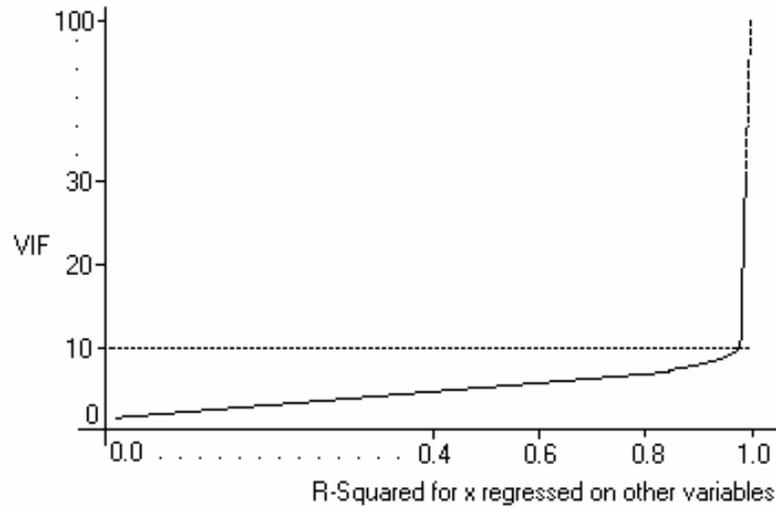


Figure 3.1 The choice of VIF value against the R-square value.

3.4.2.2 Tolerance

Tolerance is an index (set of indices) of linear dependence among the independent variables X_1, X_2, \dots, X_p in the intercept model. It is the inverse of variance inflation factors which a value of near 1 indicates the independence of the predictors while a value of close to 0 indicates the variables are multicollinear. Therefore, tolerance have a range from 0 to 1 and the closer the tolerance value is to 0, the higher the level of multicollinearity exists. It is calculated as follows :

$$(Tolerance)_j = 1 - R_j^2$$

3.4.2.3 Eigenvalues, Condition Number (CN)

The characteristic roots or eigenvalues of $\mathbf{X}^t\mathbf{X}$, say $\lambda_1, \lambda_2, \dots, \lambda_p$, can be used to measure the extent of the multicollinearity in the data. If there are one or more near-linear

dependences in the data, then one or more characteristic roots will be small. One or more small eigenvalues imply that there are near-linear dependences among the columns of \mathbf{X} . Some analyst prefer to examine the condition number of $\mathbf{X}^t\mathbf{X}$, defined as $CN = \frac{\lambda_{\max}}{\lambda_{\min}}$. This is just a measure of the spread in the eigenvalues spectrum of $\mathbf{X}^t\mathbf{X}$. Generally if the condition number is less than 100, there is no serious problem with multicollinearity. Condition number between 100 and 1000 imply moderate to strong multicollinearity, and if exceeds 1000 this indicates presence of severe multicollinearity.

3.5 Remedies of multicollinearity

Several approaches for handling multicollinearity problem have been developed such as Model Respecification, Use Additional or New Data , Principal Component Regression and Ridge Regression. Ridge regression will be discussed in details in the next chapter.

3.5.1 Model respecification

Multicollinearity is often caused by the choice of the model, such as when two highly correlated regressors are used in the regression equation. In these situations some respecification of the regression equation may lessen the impact of multicollinearity. One approach to model respecification is to redefine the regressors. For example, if X_1, X_2, X_3 are nearly linearly dependent, it may be possible to find some function such as $X = X_1X_2X_3$ or $X = (X_1 + X_2)/X_3$ that preserves the information content in the original regressors but reduces the ill conditioning. Another widely used approach to model respecification is variable elimination. That is, if X_1, X_2, X_3 are nearly linearly dependent, eliminating one regressor may be helpful in combating multicollinearity. Variable elimination is often a

highly effective technique. However, it may not provide a satisfactory solution if the regressors dropped from the model have significant explanatory power relative to the response variable y , that is eliminating regressor to reduce multicollinearity may damage the predictive power of the model. Care must be exercised in variables selection because many of the selection procedures are seriously distorted by the multicollinearity, and there is no assurance that the final model will exhibit any lesser degree of multicollinearity than was present in the original data.

3.5.2 Use additional or new data

Since multicollinearity is a sample feature, it is possible that the other sample involving the same variables collinearity may be not as serious as in the first sample. Sometimes simply increasing the size of the sample may attenuate the collinearity problem. If one uses more data, or increase the sample size, the effects of multicollinearity on the standard errors will decrease. This because the standard errors are based on the both the correlation between variables and the sample size. The larger the sample size, the smaller in the standard error. Unfortunately, collecting additional data is not always possible because of economic constraints or because of the process being studied is no longer available for sampling. Even when the additional data are available it may be inappropriate to use if the new data extend the range of the regressor variable far beyond the analyst's region of interest. Of course collecting additional data is not a viable solution to the multicollinearity problem when the multicollinearity is due to constraints on the model or on the population.

3.5.3 Principal component regression

Biased estimators of regression coefficients can be obtained by using a procedure known as principal components regression.

Consider the model in (2.4), let $\mathbf{X}^t\mathbf{X} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}^t$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ is $p \times p$ diagonal matrix of the eigenvalues of $\mathbf{X}^t\mathbf{X}$ and \mathbf{T} is $p \times p$ orthogonal matrix whose columns are the eigenvectors associated with $\lambda_1, \lambda_2, \dots, \lambda_p$. Then the above model can be written as

$$\mathbf{y} = \mathbf{X}\mathbf{T}\mathbf{T}^t\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbf{T}\mathbf{T}^t = \mathbf{I}$$

or \mathbf{y} can be written as

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

Where $\mathbf{Z} = \mathbf{X}\mathbf{T}$, and $\boldsymbol{\alpha} = \mathbf{T}^t\boldsymbol{\beta}$,

and we have

$$\mathbf{Z}^t\mathbf{Z} = \mathbf{T}^t\mathbf{X}^t\mathbf{X}\mathbf{T} = \mathbf{T}^t\mathbf{T}\mathbf{\Lambda}\mathbf{T}^t\mathbf{T} = \mathbf{\Lambda}$$

the columns of \mathbf{Z} , which define a new set of orthogonal regressors, such as

$$\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_p],$$

are referred to as principle components.

The least square estimator of $\boldsymbol{\alpha}$ is

$$\hat{\boldsymbol{\alpha}} = (\mathbf{Z}^t\mathbf{Z})^{-1}\mathbf{Z}^t\mathbf{y} = \mathbf{\Lambda}^{-1}\mathbf{Z}^t\mathbf{y}$$

And the covariance matrix of $\hat{\boldsymbol{\alpha}}$ is given by

$$V(\hat{\boldsymbol{\alpha}}) = \sigma^2(\mathbf{Z}^t\mathbf{Z})^{-1} = \sigma^2\mathbf{\Lambda}^{-1}$$

Thus a small eigenvalues $\mathbf{X}^t\mathbf{X}$ means that the variance of the corresponding regression coefficient will be large. Since $\mathbf{Z}^t\mathbf{Z} = \mathbf{\Lambda}$. We often refer to the eigenvalue λ_j as the variance of the j th principle component. If all λ_j equal to unity, the original regressors are orthogonal, while if a λ_j is exactly to zero, this implies a perfect linear relationship between the original regressors. One or more λ_j near to zero implies that multicollinearity is present.

The principle components regression approach combats multicollinearity by using less than the full set of principle components in the model. To obtain the principle components estimator, assume the regressors are arranged in order of decreasing eigenvalues, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ suppose that the last s of these eigenvalues are approximately equal to zero. In principle components regression the principal components corresponding to near zero eigenvalues are removed from the analysis and the least squares applied to the remaining components that is

$$\hat{\boldsymbol{\alpha}}_{pc} = \boldsymbol{\beta}\hat{\boldsymbol{\alpha}},$$

Where

$$\boldsymbol{\beta} = [b_1, b_2, \dots, b_p] \text{ and } b_1 = b_2 = \dots = b_{p-s} = 1 \text{ and } b_{p-s+1} = b_{p-s+2} = \dots = b_p = 0$$

Thus the principle components estimator is

$$\hat{\boldsymbol{\alpha}}_{pc} = [\hat{\alpha}_1 \ \hat{\alpha}_2 \ \dots \ \hat{\alpha}_{p-s} \ 0 \ 0 \ \dots \ 0]$$

Thus the original vector $\hat{\boldsymbol{\beta}}$ can be obtained by reverse transformation $\hat{\boldsymbol{\beta}} = \mathbf{T}\hat{\boldsymbol{\alpha}}_{pc}$ and the variance covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by

$$V(\hat{\boldsymbol{\beta}}) = \mathbf{T}V(\hat{\boldsymbol{\alpha}}_{pc})\mathbf{T}^t$$

The steps in PCR can be summarized in the following algorithm.

<p>STEP 1 : convert data to correlation form</p> $\tilde{X} = \frac{1}{\sqrt{n-1}} \left(\frac{X - \bar{X}}{\sigma_X} \right), \quad \tilde{Y} = \frac{1}{\sqrt{n-1}} \left(\frac{Y - \bar{Y}}{\sigma_Y} \right)$
<p>STEP 2 : Compute the correlation matrix for centered and scaled data</p> $\tilde{X}^t \tilde{X} = r_{XX}, \quad \tilde{X}^t \tilde{Y} = r_{XY}$
<p>STEP 3 : compute the eigenvalues, λ_i and the eigenvectors \mathbf{T} of correlation matrix</p>
<p>STEP 4 : Compute the component</p> $\mathbf{Z} = \mathbf{X}\mathbf{T}$
<p>STEP 5 : Compute eigenvalues of the components. The component associated with the smallest eigenvalue will be deleted.</p> $\hat{\alpha} = (\mathbf{Z}^t \mathbf{Z})^{-1} \mathbf{Z}^t \mathbf{y}$
<p>STEP 6 : Compute the coefficient estimate for the component after deletion</p>
<p>STEP 7 : Transform back the coefficient estimate to the original standardized Variables</p> $\hat{\beta}_{pc} = \mathbf{T} \hat{\alpha}_{pc}$
<p>STEP 7: Compute the coefficients of the natural variables</p> $b_i = \left(\frac{\sigma_Y}{\sigma_X} \right) \hat{\beta}_{i,pc}, i = 1, 2, \dots, p - r$ <p>where ; r : component eliminated.</p>
<p>STEP 8: The constant term is estimated by</p> $b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2 - \dots - b_{p-r} \bar{X}_{p-r}$

Figure3.2 : Steps in Principal Components Regression algorithm

Chapter four

Ridge Regression

Ridge regression is the modifications of the least squares method that allow biased estimators of the regression coefficients. Although it has biased estimators, it only has a small biased substantially more precise than an unbiased estimator. Therefore the estimator will be preferred since it will have a larger probability of being close to the true parameter value. In this chapter we will make an explanation of ridge regression and reviews the relevant literature on published work done recently concerning the problems of multicollinearity and for choosing the ridge parameter k when multicollinearity among the columns of the design matrix exists.

4.1 Ridge regression estimator

When the method of least squares method is applied to nonorthogonal data, very poor estimates of the regression coefficients can be obtained. The problem with the method of least squares is the requirement that $\hat{\boldsymbol{\beta}}$ be unbiased estimator of $\boldsymbol{\beta}$. To motivate the ridge estimator, we take a look at the mean squared error, $E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ of least squares estimator of $\boldsymbol{\beta}$. which can break into two parts the variance plus the squared bias

$$MSE(\hat{\boldsymbol{\beta}}) = E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = V(\hat{\boldsymbol{\beta}}) + [E(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}]^2$$

The Gauss-Markov property assures that the least squares estimator has minimum variance in the class of unbiased linear estimators. This however does not necessarily guarantee the minimum MSE.

One way to alleviate this problem is to drop the requirement that the estimator of $\boldsymbol{\beta}$ be unbiased.

Suppose that a biased estimator of $\boldsymbol{\beta}$ is found say $\widehat{\boldsymbol{\beta}}^*$ that has smaller variance than the unbiased estimator $\widehat{\boldsymbol{\beta}}$. The mean square error of $\widehat{\boldsymbol{\beta}}^*$ is defined as

$$MSE(\widehat{\boldsymbol{\beta}}^*) = E\|\widehat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}\|^2 = V(\widehat{\boldsymbol{\beta}}^*) + [E(\widehat{\boldsymbol{\beta}}^*) - \boldsymbol{\beta}]^2$$

or

$$MSE(\widehat{\boldsymbol{\beta}}^*) = V(\widehat{\boldsymbol{\beta}}^*) + [bias\ in\ \widehat{\boldsymbol{\beta}}^*]^2$$

By allowing a small amount of bias in $\widehat{\boldsymbol{\beta}}^*$, the variance of $\widehat{\boldsymbol{\beta}}^*$ can be made small such that the MSE of $\widehat{\boldsymbol{\beta}}^*$ is less than the variance of the unbiased estimator $\widehat{\boldsymbol{\beta}}$.

A number of procedures have been developed biased estimators of regression coefficients. One of these procedures is ridge regression, which is regression estimator has been introduced as an alternative to the ordinary least square estimator (OLS) in the presence of multicollinearity. This estimator originally proposed by Hoerl and Kennard (1970). Specifically the ridge estimator is defined as the solution to

$$(\mathbf{X}^t\mathbf{X} + k\mathbf{I})\widehat{\boldsymbol{\beta}}_R = \mathbf{X}^t\mathbf{y},$$

or

$$\widehat{\boldsymbol{\beta}}_R = (\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^t\mathbf{y},$$

where k is a positive number known as ridge parameter. The procedure is called ridge regression. An equivalent way is to write the ridge problem in the penalized or constrained least squares form by

Minimizing $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$, subject to $\|\boldsymbol{\beta}\|^2 \leq s$, for some constant s .

The sampling distribution are illustrated as follows (Neter, et. al., 1985).

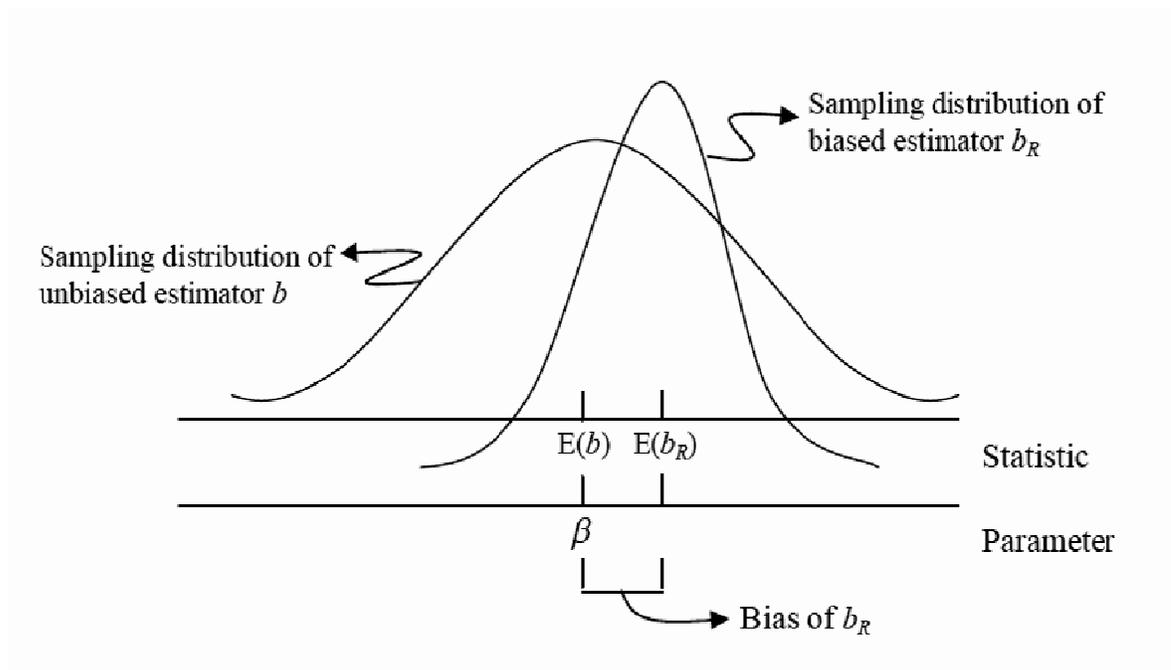


Figure 4.1 The sampling distribution of biased and unbiased estimator

Figure 4.1 illustrates the ordinary least square estimator \mathbf{b} as being unbiased but imprecise, while estimator \mathbf{b}_R is much more precise but has a small bias. Thus, the probability that \mathbf{b}_R falls near the true value $\boldsymbol{\beta}$ is much greater than that for the unbiased estimator \mathbf{b} .

The ridge solution is not invariant under scaling of the inputs. Thus we should standardize both the inputs and the response before computing the ridge estimator. With the standardized variables, the matrices $\mathbf{X}^t \mathbf{X}$ and $\mathbf{X}^t \mathbf{y}$ become

$$\mathbf{X}^t \mathbf{X} = r_{XX} \text{ and } \mathbf{X}^t \mathbf{y} = r_{Xy}$$

Where r_{XX} denotes the correlation matrix among \mathbf{X}_j and r_{Xy} denotes the correlation vector between \mathbf{y} and all \mathbf{X}_j . Hence the ridge estimator becomes

$$\widehat{\boldsymbol{\beta}}_R = (r_{XX} + k\mathbf{I})^{-1}r_{XY}$$

In the case of orthogonal predictors, the ridge estimates are just a scaled version of OLS, that is

$$\widehat{\boldsymbol{\beta}}_R = (1/1 + k)\widehat{\boldsymbol{\beta}}$$

Besides, the intercept $\boldsymbol{\beta}_0$ is automatically suppressed as 0 when working with standardized data. It is to be noted that when $k = 0$ then the ridge estimator is the least square estimator.

The ridge estimator is linear transformation of the least squares estimator since

If we denote $\mathbf{Z} = (\mathbf{I} + k(\mathbf{X}^t\mathbf{X})^{-1})^{-1}$, then

$$\widehat{\boldsymbol{\beta}}_R = \mathbf{Z}\widehat{\boldsymbol{\beta}}$$

Therefore, since $E(\widehat{\boldsymbol{\beta}}_R) = E(\mathbf{Z}\widehat{\boldsymbol{\beta}}) = \mathbf{Z}\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_R$ is a biased estimator of $\boldsymbol{\beta}$. The constant k is usually referred to the biasing parameter. The covariance matrix of $\widehat{\boldsymbol{\beta}}_R$ is

$$\mathit{cov}(\widehat{\boldsymbol{\beta}}_R) = \sigma^2\mathbf{Z}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{Z}^t$$

The total mean square error of the ridge estimator can be derived as

$$\begin{aligned} MSE(\widehat{\boldsymbol{\beta}}_R) &= V(\widehat{\boldsymbol{\beta}}_R) + [\mathit{bias\ in\ } \widehat{\boldsymbol{\beta}}_R]^2 \\ &= \mathit{tr}(\mathit{cov}(\widehat{\boldsymbol{\beta}}_R)) + \{E(\widehat{\boldsymbol{\beta}}_R) - \boldsymbol{\beta}\}^t\{E(\widehat{\boldsymbol{\beta}}_R) - \boldsymbol{\beta}\} \\ &= \sigma^2\mathit{tr}(\mathbf{Z}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{Z}^t) + \boldsymbol{\beta}^t(\mathbf{I} - \mathbf{Z})^t(\mathbf{I} - \mathbf{Z})\boldsymbol{\beta} \\ &= \sigma^2\mathit{tr}((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{Z}\mathbf{Z}^t) + k^2\boldsymbol{\beta}^t(\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-2}\boldsymbol{\beta} \\ &= \sigma^2\sum_{j=1}^p \frac{1}{\lambda_j} \cdot \frac{\lambda_j^2}{(\lambda_j + k)^2} + k^2\boldsymbol{\beta}^t(\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-2}\boldsymbol{\beta} \\ &= \sigma^2\sum_{j=1}^p \frac{\lambda_j}{(\lambda_j + k)^2} + k^2\boldsymbol{\beta}^t(\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-2}\boldsymbol{\beta} \end{aligned}$$

where $\lambda_1, \lambda_2, \dots, \lambda_p$ are the eigenvalues of $\mathbf{X}^t\mathbf{X}$. If k increases then the bias in $\widehat{\boldsymbol{\beta}}_R$ increases.

However, the variance decreases as k increases.

As k continues to increase without bound, the regression estimates all tends toward zero, because the ridge method tends to shrink the estimates of ridge coefficients toward zero. The idea of ridge regression is to pick a value of k for which the reduction in the total variance is not exceeded by the increase in the bias. If this can be done, the mean square error of the ridge estimator $\widehat{\boldsymbol{\beta}}_R$ will be less than the variance of the least square estimator $\widehat{\boldsymbol{\beta}}$.

Hoerl and Kennard (1976) proved that there exists a non zero positive value of k such that

$$MSE(\widehat{\boldsymbol{\beta}}_R) < MSE(\widehat{\boldsymbol{\beta}})$$

In other words, the ridge estimator can outperform the OLS in terms of providing a smaller MSE. Nevertheless, in practice the choice of k is yet to be determined and hence there is no guarantee that a smaller MSE always be attained by ridge regression.

The residual sum of squares of $\widehat{\boldsymbol{\beta}}_R$ is given by:

$$\begin{aligned} SS_{Res}(\widehat{\boldsymbol{\beta}}_R) &= (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_R)^t (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_R) \\ &= (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^t (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}}_R - \widehat{\boldsymbol{\beta}})^t \mathbf{X}^t \mathbf{X} (\widehat{\boldsymbol{\beta}}_R - \widehat{\boldsymbol{\beta}}) \\ &= SS_{Res}(\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}}_R - \widehat{\boldsymbol{\beta}})^t \mathbf{X}^t \mathbf{X} (\widehat{\boldsymbol{\beta}}_R - \widehat{\boldsymbol{\beta}}) \end{aligned} \quad (4.1)$$

Since the first term in the right hand side of equation (4.1) in the residual sum of squares for the least squares estimates $\widehat{\boldsymbol{\beta}}$, it is clear that as k increase, the residual sum of squares increases. Consequently, because the total sum of squares is fixed, R^2 decreases as k increases. Therefore, the ridge estimates will not necessary provide the best fit to the data, but this should not be more concerned since the interest is in obtaining a stable set of parameter estimates.

4.2 Selection of variables in ridge regression by ridge trace

Variable selection procedure often do not perform well when the predictor variables are highly correlated Marguardt and Snee (1970) point out that when the data is highly multicollinear, the maximum variance inflation completely destabilizes all the criteria obtained from the least squares estimates. Hoerl and Kennard suggest that the ridge trace can be used as a guide for variable selection. They propose the following procedure for eliminating predictor variables from the full model.

1. Eliminate predictor variables that are stable but have small predicting power that is those with small standardized regression coefficient.
2. Eliminate predictor variables with unstable coefficients that do not hold their predicting power because the coefficients tend to zero as k increases.
3. Eliminate one or more of the remaining predictor variables that have small coefficients. The subset of remaining predictor variable is used in the final model.

4.3 General ridge regression.

In general ridge regression p ridge parameters have to be determined, but in ridge regression we need to find one ridge parameter. To discuss the properites of genrerall ridge regression estimator we usually tansform the linear regression model (2.4) to a canonical form. It is clear that for $p \times p$ positive definite matrix $\mathbf{X}^t\mathbf{X}$, there exists a $p \times p$ orthogonal matrix \mathbf{T} such that $\mathbf{T}^t \mathbf{X}^t\mathbf{X}\mathbf{T} = \mathbf{\Lambda}$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ and $\lambda_1 \geq \lambda_1 \geq \dots \geq \lambda_p$ are the orderd eigenvalues of $\mathbf{X}^t\mathbf{X}$ matrix. We may write (2.4) as

$$\mathbf{y} = \mathbf{H}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

The ordinary least squares estimator of α is

$$\hat{\alpha} = \Lambda^{-1}c \quad (4.2)$$

Where, $H = XT$, $\alpha = T^t\beta$ and $c = H^ty$

In scalar notation we can write (4.2) as

$$\hat{\alpha}_i = \frac{c_i}{\lambda_i}, i = 1, 2, \dots, p$$

And so we can write the ridge estimator as

$$\hat{\alpha}^R = (\Lambda + kI)^{-1}H^ty \quad (4.3)$$

In scalar notation we can write (4.3) as

$$\hat{\alpha}_i^R = \frac{c_i}{\lambda_i + k}, i = 1, 2, \dots, p$$

In this study two type of general ridge regression will be considered

4.3.1 General ridge regression I.

The general ridge regression can be written as

$$\hat{\alpha}^{GR} = (\Lambda + kI)^{-1}H^ty \quad (4.4)$$

Where $K = \text{diag}(k_1, k_2, \dots, k_p)$ and k_i is a positive number for each $i = 1, 2, \dots, p$, equation

(4.4) is called the general form of ridge regression (GR) which is proposed by Hoerl and

Kennard, 1970; in scalar notation (4.4) can be written as

$$\hat{\alpha}_i^{GR} = \frac{c_i}{\lambda_i + k_i}, i = 1, 2, \dots, p$$

In ridge regression all eigenvalues of $\mathbf{X}^t\mathbf{X}$ are treated equally, while in general ridge regression, the determination of p ridge parameters k_1, k_2, \dots, k_p is required.

It follows from Hoerl and Kennard, 1970; that the value of k_i which minimizes the $MSE(\hat{\boldsymbol{\alpha}}^{GR})$, where

$$MSE(\hat{\boldsymbol{\alpha}}^{GR}) = \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k_i)^2} + \sum_{i=1}^p \frac{k_i^2 \alpha_i^2}{(\lambda_i + k_i)^2} \quad (4.5)$$

is

$$k_i = \frac{\sigma^2}{\alpha_i^2} \quad (4.6)$$

Where σ^2 is the error variance of model (2.3) and α_i is the i^{th} element of $\boldsymbol{\alpha}$

Equation (4.6) gives a value of k_i that fully depends on the unknown σ^2 and α_i and must be estimated from the observed data. Hoerl and Kennard, 1970; suggest the replacement of σ^2 and α_i by their corresponding unbiased estimators, that is

$$\hat{k}_i = \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}$$

Where $\hat{\sigma}^2 = \sum_{i=1}^n \varepsilon_i^2 / n - p$ is the residual mean square estimate, which is unbiased estimator of σ^2 , and $\hat{\alpha}_i$ is the element of $\hat{\boldsymbol{\alpha}}$, which is unbiased estimator $\boldsymbol{\alpha}$. They found that the best method for achieving a better estimate $\hat{\boldsymbol{\alpha}}_R$ is to use $k_i = k$ for each i and they suggest k to be \hat{k}_{HK} where

$$\hat{k}_{HK} = \frac{\hat{\sigma}^2}{\max(\hat{\alpha}_i)}$$

If σ^2 and α are known, then \hat{k}_{HK} is sufficient to give ridge estimators having smaller mean square error than the ordinary least square estimators.

4.3.2 Superiority of the GRR $\hat{\alpha}^{GR}$ over OLS $\hat{\alpha}$

J.S. Chawla, (1989); gave a sufficient condition for k_i such that the general ridge regression, $\hat{\alpha}^{GR}$ given in (4.4) is better than the ordinary least square $\hat{\alpha}$ given in (4.2) relative to the mean square error.

In the following theorem a sufficient condition for k_i will be considered. The proof of the theorem requires the following two lemmas:

Lemma (4.1)

$$MSE(\hat{\alpha}^{GR}) = \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k_i)^2} + \sum_{i=1}^p \frac{k_i^2 \alpha_i^2}{(\lambda_i + k_i)^2}$$

Proof.

$$\begin{aligned} MSE(\hat{\alpha}^{GR}) &= E(\hat{\alpha}^{GR} - \alpha)^t (\hat{\alpha}^{GR} - \alpha) = E(\mathbf{Z}\hat{\alpha} - \alpha)^t (\mathbf{Z}\hat{\alpha} - \alpha) \\ &= E(\mathbf{Z}\hat{\alpha} - \mathbf{Z}\alpha)^t (\mathbf{Z}\hat{\alpha} - \mathbf{Z}\alpha) + (\mathbf{Z}\alpha - \alpha)^t (\mathbf{Z}\alpha - \alpha) \end{aligned}$$

where

$$\mathbf{Z} = (\mathbf{I} + \mathbf{A}^{-1}\mathbf{K})^{-1} \text{ and } \hat{\alpha}^{GR} = \mathbf{Z}\hat{\alpha}$$

Now,

$$\begin{aligned} E(\mathbf{Z}\hat{\alpha} - \mathbf{Z}\alpha)^t (\mathbf{Z}\hat{\alpha} - \mathbf{Z}\alpha) &= \text{trace}[E(\mathbf{Z}\hat{\alpha} - \mathbf{Z}\alpha)(\mathbf{Z}\hat{\alpha} - \mathbf{Z}\alpha)^t] \\ &= \text{trace}[\mathbf{Z}E(\hat{\alpha} - \alpha)(\hat{\alpha} - \alpha)^t \mathbf{Z}^t] \\ &= \text{trace}[\mathbf{Z}\sigma^2 \mathbf{A}^{-1} \mathbf{Z}^t] \end{aligned}$$

Thus we have

$$\begin{aligned}
MSE(\hat{\alpha}^{GR}) &= \sigma^2 \text{trace}[\mathbf{Z}\mathbf{\Lambda}^{-1}\mathbf{Z}^t] + \boldsymbol{\alpha}^t(\mathbf{Z} - \mathbf{I})^t(\mathbf{Z} - \mathbf{I})\boldsymbol{\alpha} \\
&= \sigma^2 \text{trace}[(\mathbf{I} + \mathbf{\Lambda}^{-1}\mathbf{K})^{-1}\mathbf{\Lambda}^{-1}(\mathbf{I} + \mathbf{\Lambda}^{-1}\mathbf{K})^{-1}] \\
&\quad + \boldsymbol{\alpha}^t[(\mathbf{I} + \mathbf{\Lambda}^{-1}\mathbf{K})^{-1} - \mathbf{I}]^t[(\mathbf{I} + \mathbf{\Lambda}^{-1}\mathbf{K})^{-1} - \mathbf{I}]\boldsymbol{\alpha} \\
&= \sigma^2 \sum_{i=1}^p [(1/\lambda_i)(1 + k_i/\lambda_i)^{-2}] + \sum_{i=1}^p [(1 + k_i/\lambda_i)^{-1} - 1]^2 \alpha_i^2 \\
&= \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k_i)^2} + \sum_{i=1}^p \frac{k_i^2 \alpha_i^2}{(\lambda_i + k_i)^2}
\end{aligned}$$

Lemma (4.2)

$$\sigma^2 \frac{\lambda_i}{(\lambda_i + k)^2} + \frac{k^2 \alpha_i^2}{(\lambda_i + k)^2} \text{ is a monotonically decreasing function if } 0 \leq k \leq \sigma^2 / \alpha_i^2.$$

Proof. Let $f_i(k) = \sigma^2 \frac{\lambda_i}{(\lambda_i + k)^2} + \frac{k^2 \alpha_i^2}{(\lambda_i + k)^2}$

Therefore,

$$\frac{df_i(k)}{dk} = (-2\sigma^2 \lambda_i + 2\alpha_i^2 \lambda_i k) / (\lambda_i + k)^3.$$

If

$$k \leq \sigma^2 / \alpha_i^2, \text{ then } \frac{df_i(k)}{dk} \leq 0$$

That is, $\sigma^2 \frac{\lambda_i}{(\lambda_i + k)^2} + \frac{k^2 \alpha_i^2}{(\lambda_i + k)^2}$ is a monotonically decreasing.

Theorem (4.1)

$$MSE(\hat{\alpha}^{GR}) < MSE(\hat{\alpha}) \text{ if the largest } k_i < \Omega; \text{ where } \Omega = \min \{\sigma^2 / \alpha_i^2 \mid i = 1, 2, \dots, p\}$$

Proof.

From lemma (4.2)

$$f_i(k) < f_i(0)$$

or

$$\sigma^2 \frac{\lambda_i}{(\lambda_i+k)^2} + \frac{k^2 \alpha_i^2}{(\lambda_i+k)^2} < \sigma^2 / \lambda_i$$

If $0 \leq k \leq \sigma^2 / \alpha_i^2$. Hence

$$\sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k_i)^2} + \sum_{i=1}^p \frac{k_i^2 \alpha_i^2}{(\lambda_i + k_i)^2} < \sum_{i=1}^p \sigma^2 / \lambda_i$$

If largest $k_i < \min \{\sigma^2 / \alpha_i^2 \mid i = 1, 2, \dots, p\}$. Equivalently, $MSE(\hat{\alpha}^{GR}) < MSE(\hat{\alpha})$, if the largest $k_i < \Omega$.

4.3.3 General ridge regression(II).

Farebrother, (1978); proposed an estimator of β in model (2.4) given by

$$\hat{\alpha}^* = (\mathbf{X}^t \mathbf{X} + k\mathbf{A})^{-1} \mathbf{X}^t \mathbf{y}, \quad (4.7)$$

Where k is a positive number and \mathbf{A} is $p \times p$ positive semi-definite matrix.

If \mathbf{b} is a biased estimator of β , then the $p \times p$ matrix of mean square error of \mathbf{b} is defined as

$$Mtx \mathbf{MSE}(\mathbf{b}) = E(\mathbf{b} - \beta)(\mathbf{b} - \beta)^t$$

Chawla, 1988; found that

$$Mtx \mathbf{MSE}(\hat{\alpha}^*) = (\mathbf{X}^t \mathbf{X} + k\mathbf{A})^{-1} [\sigma^2 \mathbf{X}^t \mathbf{X} + k^2 \mathbf{A} \beta \beta^t \mathbf{A}] (\mathbf{X}^t \mathbf{X} + k\mathbf{A})^{-1} \quad (4.8)$$

If \mathbf{b}_1 and \mathbf{b}_2 are two competing estimators of β and

$$\Delta = Mtx \mathbf{MSE}(\mathbf{b}_2) - Mtx \mathbf{MSE}(\mathbf{b}_1)$$

is positive definite, then \mathbf{b}_1 is preferred to \mathbf{b}_2 .

4.3.4 Superiority of $\hat{\alpha}^*$ over the ordinary least square $\hat{\beta}$

General ridge estimator excels the least square estimator under a necessary and sufficient condition, using the matrix mean square error criterion.

The following theorem gives these necessary and sufficient conditions, the proof of this theorem requires the following lemma.

Lemma (4.3): Let \mathbf{R} be a $p \times m$ matrix of rank m such that $\mathbf{A} = \mathbf{R}\mathbf{R}^t$, then

$$\begin{aligned} \mathbf{S} &= \text{Mtx MSE}(\hat{\beta}) - \text{Mtx MSE}(\hat{\alpha}^*) \\ &= k^2\sigma^2(\mathbf{X}^t\mathbf{X} + k\mathbf{A})^{-1}\mathbf{R}\mathbf{Q}\mathbf{R}^t(\mathbf{X}^t\mathbf{X} + k\mathbf{A})^{-1} \end{aligned} \quad (4.9)$$

Where

$$\mathbf{Q} = (2/k)\mathbf{I}_m + \mathbf{R}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{R} - (1/\sigma^2)\mathbf{R}^t\boldsymbol{\beta}\boldsymbol{\beta}^t\mathbf{R} \quad (4.10)$$

Proof.

Substitute $\text{Mtx MSE}(\hat{\beta}) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$ and equation (4.8) into (4.9) we get

$$\begin{aligned} \mathbf{S} &= \sigma^2(\mathbf{X}^t\mathbf{X})^{-1} - (\mathbf{X}^t\mathbf{X} + k\mathbf{A})^{-1}[\sigma^2\mathbf{X}^t\mathbf{X} + k^2\mathbf{A}\boldsymbol{\beta}\boldsymbol{\beta}^t\mathbf{A}](\mathbf{X}^t\mathbf{X} + k\mathbf{A})^{-1} \\ &= (\mathbf{X}^t\mathbf{X} + k\mathbf{A})^{-1}[\sigma^2(\mathbf{X}^t\mathbf{X} + k\mathbf{A})(\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\mathbf{X} + k\mathbf{A}) \\ &\quad - \sigma^2(\mathbf{X}^t\mathbf{X}) - k^2\mathbf{A}\boldsymbol{\beta}\boldsymbol{\beta}^t\mathbf{A}] \times (\mathbf{X}^t\mathbf{X} + k\mathbf{A})^{-1} \\ &= (\mathbf{X}^t\mathbf{X} + k\mathbf{A})^{-1}[2\sigma^2k\mathbf{A} + \sigma^2k^2\mathbf{A}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{A} - k^2\mathbf{A}\boldsymbol{\beta}\boldsymbol{\beta}^t\mathbf{A}](\mathbf{X}^t\mathbf{X} + k\mathbf{A})^{-1} \\ &= \sigma^2k^2(\mathbf{X}^t\mathbf{X} + k\mathbf{A})^{-1}\mathbf{R}[(2/k)\mathbf{I}_m + \mathbf{R}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{R} \\ &\quad - (1/\sigma^2)\mathbf{R}^t\boldsymbol{\beta}\boldsymbol{\beta}^t\mathbf{R}]\mathbf{R}^t(\mathbf{X}^t\mathbf{X} + k\mathbf{A})^{-1} \\ &= k^2\sigma^2(\mathbf{X}^t\mathbf{X} + k\mathbf{A})^{-1}\mathbf{R}\mathbf{Q}\mathbf{R}^t(\mathbf{X}^t\mathbf{X} + k\mathbf{A})^{-1}. \quad \blacksquare \end{aligned}$$

Theorem (4.2): A necessary and sufficient condition for

$$\mathbf{S} = \text{Mtx MSE}(\hat{\beta}) - \text{Mtx MSE}(\hat{\alpha}^*)$$

to be positive definite is

$$0 < k < 2/|\emptyset|$$

Where \emptyset is the smallest negative eigenvalue of

$$\mathbf{R}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{R} - (1/\sigma^2)\mathbf{R}^t\boldsymbol{\beta}\boldsymbol{\beta}^t\mathbf{R} \quad (4.11)$$

If all the eigenvalues of (4.11) are nonnegative, then \mathbf{S} is positive definite for all values of $k > 0$.

Proof.

\mathbf{S} is positive definite if and only if (4.10) is positive definite. Let $\tau_1, \tau_2, \dots, \tau_m$ be the eigenvalues of (4.11) therefore the eigenvalues of (4.10) are

$$(2/k) + \tau_1, (2/k) + \tau_2, \dots, (2/k) + \tau_m.$$

If all $\tau_i \geq 0$, $i = 1, 2, \dots, m$, then (4.10) is positive definite for all values $k > 0$. If some $\tau_i < 0$, then \emptyset is the least value of τ_i , $i = 1, 2, \dots, m$, therefore (4.10) is positive definite if and only if $(2/k) + \emptyset > 0$. This equivalent to $0 < k < 2/|\emptyset|$. ■

4.4 Ridge parameter k

Hoerl and Kennard (1976) have suggested that an appropriate value of k may be determined by the ridge trace. The ridge trace is a plot of the elements of $\widehat{\boldsymbol{\beta}}_R$ versus k for values of k usually in the interval $[0,1]$. If the multicollinearity is severe, the instability in the regression coefficients will be obvious from the ridge trace. As k is increased, some of the ridge estimates will vary dramatically. At some value of k , the ridge estimates $\widehat{\boldsymbol{\beta}}_R$ will

stabilize. The objective is to select a reasonable small value of k at which the ridge estimates $\hat{\beta}_R$ are stable.

Several author have proposed several procedures for choosing the value of k . Hoerl and Kennard (1970a) proposed $k_{HKa} = \frac{\hat{\sigma}^2}{\hat{\beta}_{max}^2}$ to estimate the ridge parameter k , also they have suggested in (1975) that an appropriate choice of k is $k = p\hat{\sigma}^2/\hat{\beta}^t\hat{\beta}$, where $\hat{\beta}$ and $\hat{\sigma}^2$ are found by least squares solution and p is the number of parameter.

Hoerl and Kennard recommended $k_{HK} = \frac{p\hat{\sigma}^2}{\hat{\beta}^t\hat{\beta}}$ as general rule where the parameters are estimated from the full equation least squares fit. Their studies suggest that the resulting ridge estimator yields coefficient estimates with smaller means squared error than the obtained from least squares. In a latter paper Hoerl and Kennard (1975) suggest an iterative procedure where $k = \frac{p\hat{\sigma}^2}{\hat{\beta}_i^t\hat{\beta}_i}$ where $\hat{\beta}_i = \hat{\beta}_R(k_i)$. Farebrother (1975) suggested $k = \frac{\hat{\sigma}^2}{\hat{\beta}^t\hat{\beta}}$, which for the Gonman-Toman data, yield $k = 0.003$ with this formula. Marquardt and Snee (1970) suggested value of k for which the maximum variance inflation factor is between one and ten. Mallows (1973) extended the concept of C_p – plots to C_k –plots, which may be used to determine k Specifically, he suggested plotting C_k , versus V_k where

$$C_k = (RRS_k/\hat{\sigma}^2) - n + 2 + Tr(\mathbf{X}\mathbf{L})$$

$$V_k = 1 + Tr(\mathbf{X}^t\mathbf{X}\mathbf{L}\mathbf{L}^t)$$

And

$$\mathbf{L} = (\mathbf{X}^t\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^t$$

Here RRS_k is the residual sum of squares as a function of k the suggestion is to choose k to minimize C_k . And other several methods for estimating k have been proposed by Galarneau, 1975; Lawless and Wang, 1976; Hocking et al. 1976; Wichern and Churchill, 1978; Nordberg, 1982; Saleh and Kibria, 1993; Singh and Tracy, 1999; Wenchekeo, 2000; Kibria, 2003; Alkhamisi et al., 2006; and Alkhamisi and Shukur, 2007; Alkhamisi and Shukur, 2007; proposed some new estimators by adding $1/\lambda_{max}$ to some will known estimator, where λ_{max} is the largest eigenvalue of $\mathbf{X}^t \mathbf{X}$.

Khalf and Shukur, 2005; suggested an estimator based on \hat{k}_{HKa} named as \hat{k}_{KS} , where

$$\hat{k}_{KS} = \frac{\lambda_{max} \hat{\sigma}^2}{\lambda_{max} \hat{\beta}_{max}^2 + (n - p) \hat{\sigma}^2}$$

Hocking et al., 1976; suggest an estimator depend on (4.2) estimator named as \hat{k}_{HSL} , or (HSL) for k

Where

$$\hat{k}_{HSL} = \hat{\sigma}^2 \frac{\sum_{i=1}^p (\lambda_i \hat{\alpha}_i)^2}{(\sum_{i=1}^p \lambda_i \hat{\alpha}_i^2)^2}$$

Mahdi ALkamisi, Ghadban Khalaf and Ghazi Shukur 2006; proposed Some Modifications for Choosing Ridge Parameters as follows:

$$\hat{k}_{KSHK} = \max \left(\frac{\hat{\sigma}^2}{\hat{\beta}_i} \right), i = 1, 2, \dots, p$$

$$\hat{k}_{KS MAX} = \max \left(\frac{\lambda_i \hat{\sigma}^2}{\lambda_i \hat{\beta}_i^2 + (n - p) \hat{\sigma}^2} \right)$$

$$\hat{k}_{KSMED} = \text{median} \left(\frac{\lambda_i \hat{\sigma}^2}{\lambda_i \hat{\beta}_i^2 + (n-p) \hat{\sigma}^2} \right), i = 1, 2, \dots, p$$

$$\hat{k}_{KSARTH} = \frac{1}{p} \sum_{i=1}^p \left(\frac{\lambda_i \hat{\sigma}^2}{\lambda_i \hat{\beta}_i^2 + (n-p) \hat{\sigma}^2} \right)$$

M. A. ALkhamis and G. Shukur, 2007 presented a new method based on k_{HK}

These estimators is presented as follows:

$$\hat{k}_{AS} = \frac{\hat{\sigma}^2}{\hat{\beta}_{max}^2} + \frac{1}{\lambda_{max}}$$

$$\hat{k}_{NHKB} = \frac{p \hat{\sigma}^2}{\hat{\beta}^t \hat{\beta}} + \frac{1}{\lambda_{max}}$$

$$\hat{k}_{NAS} = \text{Max} \left(\frac{\hat{\sigma}^2}{\hat{\beta}_i^2} + \frac{1}{\lambda_i} \right), i = 1, 2, \dots, p$$

$$\hat{k}_{ARITH} = \frac{1}{p} \sum_{i=1}^p \left(\frac{\hat{\sigma}^2}{\hat{\beta}_i^2} + \frac{1}{\lambda_i} \right)$$

$$\hat{k}_{NMED} = \text{Median} \left(\frac{\hat{\sigma}^2}{\hat{\beta}_i^2} + \frac{1}{\lambda_i} \right), i = 1, 2, \dots, p$$

$$\hat{k}_{NLW} = \frac{p \hat{\sigma}^2}{\sum_{i=1}^p \lambda_i \hat{\beta}_i^2} + \frac{1}{\lambda_{max}}$$

Yazid M. AL-Hassan, 14 December 2010; apply the modification mentioned in Alkhamisi and Shukur, 2007; to the estimator proposed by Hocking et al. 1976, \hat{k}_{HSL} , to obtain new estimator named \hat{k}_{NHSL} , or (NHSL)

Where

$$\begin{aligned}\hat{k}_{NHSL} &= \hat{\sigma}^2 \frac{\lambda_{max} \sum_{i=1}^p (\lambda_i \hat{\alpha}_i)^2 + (\sum_{i=1}^p \lambda_i \hat{\alpha}_i^2)^2}{\lambda_{max} (\sum_{i=1}^p \lambda_i \hat{\alpha}_i^2)^2} \\ &= \hat{\sigma}^2 \frac{\sum_{i=1}^p (\lambda_i \hat{\alpha}_i)^2}{(\sum_{i=1}^p \lambda_i \hat{\alpha}_i^2)^2} + \frac{1}{\lambda_{max}} = \hat{k}_{HSL} + \frac{1}{\lambda_{max}}\end{aligned}$$

Since $\frac{1}{\lambda_{max}} \geq 0$, \hat{k}_{NHSL} is greater than \hat{k}_{HSL} .

Yazid M. AL-Hassan, 2010; used Monte Carlo simulation to investigate the properties of OLS, HK, HSL and NHSL. And he made a comparison between these estimators based on the MSE criterion. That is, he compared OLS, HK and HSL estimators with NHSL. He found that his modified estimator NHSL uniformly dominates the other estimators OLS, HK, and HSL.

In this study we will make a comparison between the OLS and other approach for choosing the ridge parameter k , these approach is listed in the following table

Table 4.1 Ridge parameters which we made a comparison between them.

Name	k
HKa	$\frac{\hat{\sigma}^2}{\hat{\beta}_{max}^2}$
KS	$\frac{\lambda_{max} \hat{\sigma}^2}{\lambda_{max} \hat{\beta}_{max}^2 + (n-p) \hat{\sigma}^2}$
HK	$\frac{p \hat{\sigma}^2}{\hat{\beta}^t \hat{\beta}}$
FK	$\frac{\hat{\sigma}^2}{\hat{\beta}^t \hat{\beta}}$

Chapter five

Applications

The early stages of this thesis are discussed how the chosen methods perform in regression analysis to handle multicollinearity problems. In the last stage of this thesis specially in this chapter we will evaluate the performance of ridge regression approaches by conducting a simulation studies to examine the feasibility and the properties of OLS, HKa, KS , HK and KF . We investigate how will the regression parameters can be estimated in terms of bias and converge rate, and then a comparison is made based on the MSE criterion. Also we study how the following factors affect the performance of these approaches: the sample size n , the number of regressors p , and degree of correlation. Moreover, a real data set also will be examined.

5.1 Simulated data

5.1.1 Generating Simulated Data Sets

The more number of regressors involved, the more chances to have multicollinearity problems in the analysis. A number of factors can affect the properties of OLS and the ridge parameters such as the sample size n , degree of correlation between the explanatory variables r , and the number of regressors p . The numbers of independent variables and the number of observations is generated randomly to test the performance of ridge parameters. The different degree of correlation between the variables included in the model has been used. We put these values equals to 0.7, 0.8, 0.9, 0.95, and 0.998. These values will cover a wide range of moderate and strong correlation between the variables. All these values show

that the correlations between all variables within different sets of regressors are very high. So, multicollinearity problems exist in the simulated data.

The response variable that is considered in this simulation study is univariate. The regression condition for this study is shown in Table 5.1.

Table 5.1 Factors and levels for the simulated data sets

Factors	Levels
Number of regressor variables	2, 5,6,10,12, 15,20,30,40, 60, 70
Number of observations	15, 30, 50, 80, 100
High correlation between regressors	0.998, 0.95, 0.9, 0.8, 0.7

The observations X_i were generated according to the model

$$y_i = X_{ip}\beta_p + \varepsilon_i, i = 1, 2, \dots, n$$

Where X_{ip} is generated from $N(0, \Sigma)$ distribution as shown in tables 5.2, 5.3 and 5.4. For the purpose of obtaining collinearity in each set of data, the X_p , were generated according to $X_p = X_1 + \Delta_p$, and the columns of the noise matrix Δ_p are independently distributed according to $N(0, c)$, where values of c determine the correlation between the regressors, we note that as c increases the correlation between the regressors decreases. Three different groups of data were generated as shown in tables 5.2, 5.3, 5.4.

For each set of the simulated data, the distribution of the random error for every set of n observations is $N(0,1)$, the number of replications, m is set to 1000 data sets. The value of ($m = 1000$) is chosen because it is enough to show a consistent results for each generated data sets.

Table 5.2 Group one of simulated data

p	n	X_{ip}
2	15, 30, 50, 80	$X_1 = N(0,1)$ $X_p = X_1 + N(0,.1), p = 2$
5	30	$X_1 = N(0,1)$ $X_p = X_1 + N(0,.1), p = 2, 3, 4, 5$
10	15, 30, 50, 80	$X_1 = N(0,1)$ $X_p = X_1 + N(0,.1), p = 2, 3, \dots, 10$
15	30	$X_1 = N(0,1)$ $X_p = X_1 + N(0,.1), p = 2, 3, \dots, 15$
20	50, 80	$X_1 = N(0,1)$ $X_p = X_1 + N(0,.1), p = 2, 3, \dots, 20$
30	50	$X_1 = N(0,1)$ $X_p = X_1 + N(0,.1), p = 2, 3, \dots, 30$
40	80	$X_1 = N(0,1)$ $X_p = X_1 + N(0,.1), p = 2, 3, \dots, 40$
60	80	$X_1 = N(0,1)$ $X_p = X_1 + N(0,.1), p = 2, 3, \dots, 60$
$y = X_1 + X_2 + \dots + X_p + N(0,1)$		

Table 5.3 Group two of simulated data

n	p	X_{ip}
100	6	$X_1 = N(0,1)$ $X_p = X_1 + N(0, c), p = 2,3 \dots,6$ $c = 0.5, 1$
	12	$X_1 = N(0,1)$ $X_p = X_1 + N(0, c), p = 2,3, \dots,12$ $c = 0.5, 1$
	20	$X_1 = N(0,1)$ $X_p = X_1 + N(0, c), p = 2,3, \dots,20$ $c = 0.5, 1$
	50	$X_1 = N(0,1)$ $X_p = X_1 + N(0, c), p = 2,3, \dots,50$ $c = 0.5, 1$
	70	$X_1 = N(0,1)$ $X_p = X_1 + N(0, c), p = 2,3, \dots,70$ $c = 0.5, 1$
$y = X_1 + X_2 + \dots + X_p + N(0,1)$		

Table 5.4 Group three of simulated data

p	n	X_{ip}
2	15,30,50,80,100	$X_1 = N(0,1)$ $X_p = X_1 + N(0, c), p = 2$ $c = 0.4, 0.5, 0.9, 1$
$y = X_1 + X_2 + \dots + X_p + N(0,1)$		

5.1.2 Performance of Ridge Regression to simulated data

Afifi and Clark (1984) stated that when two or more variables are highly correlated (greater than 0.95), it may be simplest to use only one of them, since one variable conveys essentially all of the information contained in the other. However, Wesolowsky (1976) stated that when an independent variable that is correlated with others in the regression is not included and the regression parameter of this variable is not zero, the remaining coefficients will be biased estimators, but even if the omitted variable is not correlated (in the sample) with the remaining variables the estimators for the variances of the remaining coefficients $S^2_{b_i}$, will tend to be too large. This occurs because the 'explanatory' power of the missing variable is removed, causing a larger sum of squared residuals, which, in turn, swells the variances of the regression coefficients. As a results, it becomes more difficult to show the significance of coefficients. Thus, in this study Ridge regression, will be used to handle these problems rather than using omitted variables approach.

Figure 5.1 and 5.2 illustrates how the performances of ridge regression.

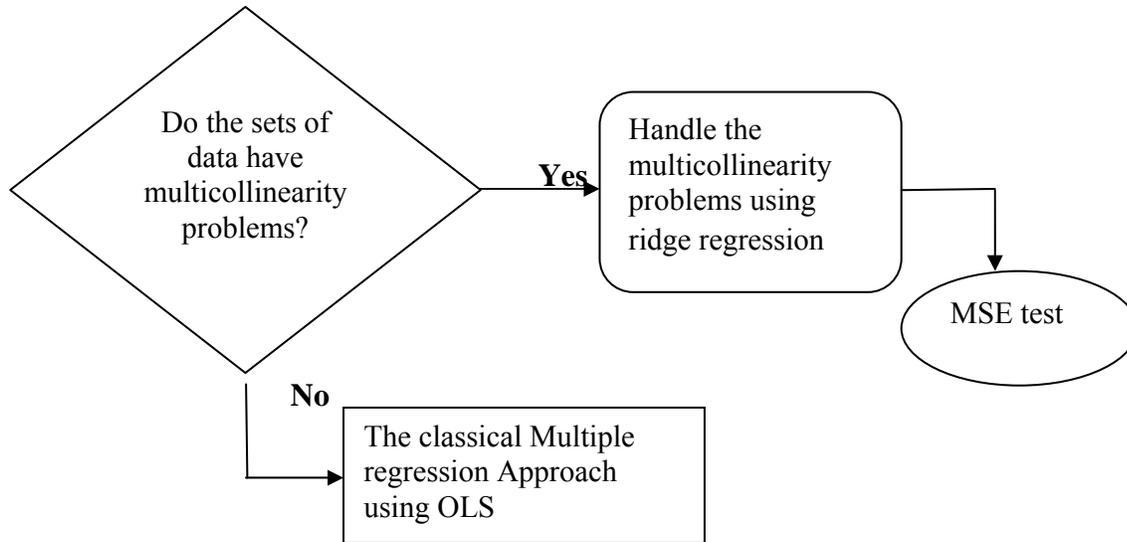


Figure 5.1 Flowchart summarizing performance of RR.

<p>STEP 1 : Convert data into correlation form</p> $\frac{1}{\sqrt{n-1}} \left(\frac{X - \bar{X}}{\sigma_X} \right) = \tilde{X} \qquad \frac{1}{\sqrt{n-1}} \left(\frac{Y - \bar{Y}}{\sigma_Y} \right) = \tilde{Y}$
<p>STEP 3 : Compute the ridge parameter k for the values</p> $\hat{k}_{HKa}, \hat{k}_{KS}, \hat{k}_{HK}, \hat{k}_{KF}$
<p>STEP 4 : Compute the ridge regression estimators for the values of k</p> $\tilde{b} = (r_{XX} + kI)^{-1} r_{XY}$
<p>STEP 5: Compute the $\mathbf{MSE}(\tilde{b}) = \frac{1}{1000} \sum_{r=1}^{1000} (\tilde{b} - b)^t (\tilde{b} - b)$ for each value of k.</p>
<p>STEP 6 : Choose the model with least MSE of \tilde{b}</p>

Figure 5.2 : Steps in Ridge Regression algorithm used in this thesis

5.1.3 Simulation results

Our primary interest lies in the investigating the properties of well known approach to minimize the MSE, In this section we present the results of simulated data for each group of the three groups concerning the properties of these approach for choosing the ridge parameter k , when multicollinearity among the columns of the design matrix exists. Our primary interest lies in comparing the MSEs of these methods for choosing the ridge parameter k that are used in this study, i.e., the HKa, KS, HK, and KF. To compare the performances of the considered estimators, we calculate the MSEs of each one. We consider the estimator that leads to the minimum MSE to be the best. It is worth mentioning here in that we used the *Matlab 10* program to simulate the data and to do all calculations that were made in this thesis. The program that we are based on to generate simulated data is sited in appendix A.

The problem of multicollinearity can also be seen through correlation matrix between regressors. The value close to 1 shows a strong relation among the regressors. The correlation results of group one are shown in Tables 5.5 –5.12 specifically for $n = 15, 30, 50, 80$ observations and for $p = 2, 5, 10$. For group one of simulations data The smallest and the highest correlation values are vary between 0.98 and 0.9988. the higher correlation of group two are shown in tables 5.15. The estimated MSEs for the three groups of simulation data are shown in tables 5.14, 5.15, 5.16.

Table 5.5 The value of correlation for $p = 10, n = 15$

$p = 10$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1									
X_2	0.99595	1								
X_3	0.99581	0.99473	1							
X_4	0.99808	0.99646	0.9979	1						
X_5	0.99669	0.99416	0.99179	0.99448	1					
X_6	0.99465	0.98792	0.98955	0.99125	0.99323	1				
X_7	0.99415	0.99313	0.98877	0.9943	0.99082	0.98794	1			
X_8	0.99719	0.9951	0.99386	0.99733	0.99238	0.9851	0.99215	1		
X_9	0.99695	0.99488	0.99139	0.99462	0.9925	0.98692	0.99461	0.99619	1	
X_{10}	0.99645	0.99258	0.9897	0.99466	0.9927	0.98673	0.99111	0.99696	0.99384	1

Table 5.6 The value of correlation for $p = 10, n = 30$

$p = 10$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1									
X_2	0.99548	1								
X_3	0.99612	0.99497	1							
X_4	0.99548	0.99011	0.99164	1						
X_5	0.99672	0.99582	0.9923	0.99293	1					
X_6	0.99668	0.99297	0.99242	0.99355	0.9936	1				
X_7	0.99506	0.99152	0.99254	0.99128	0.99093	0.99051	1			
X_8	0.99606	0.98997	0.99212	0.99262	0.99352	0.99201	0.99234	1		
X_9	0.99529	0.99254	0.98888	0.99006	0.99582	0.98986	0.99177	0.99198	1	
X_{10}	0.99404	0.991	0.99078	0.98625	0.9933	0.99228	0.99026	0.99096	0.99052	1

Table 5.7 The value of correlation for $p = 10, n = 50$

$p = 10$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1									
X_2	0.99633	1								
X_3	0.99471	0.98884	1							
X_4	0.99572	0.99183	0.99179	1						
X_5	0.99642	0.99302	0.99001	0.99202	1					
X_6	0.99416	0.99052	0.98914	0.99073	0.99188	1				
X_7	0.99505	0.99233	0.99053	0.99167	0.98995	0.98745	1			
X_8	0.99461	0.9915	0.99053	0.99005	0.99244	0.99137	0.99006	1		
X_9	0.99584	0.99187	0.991	0.99353	0.99239	0.99057	0.98888	0.98895	1	
X_{10}	0.9955	0.99195	0.99052	0.99164	0.99139	0.99145	0.98843	0.98712	0.99294	1

Table 5.8 The value of correlation for $p = 10, n = 80$

$p = 10$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1									
X_2	0.99583	1								
X_3	0.99431	0.99122	1							
X_4	0.9962	0.99284	0.98995	1						
X_5	0.9958	0.99183	0.98803	0.99355	1					
X_6	0.99558	0.99049	0.99217	0.992	0.98928	1				
X_7	0.99649	0.9947	0.99056	0.99321	0.99326	0.99176	1			
X_8	0.99468	0.99056	0.98981	0.99178	0.99128	0.98966	0.99134	1		
X_9	0.99603	0.99004	0.98925	0.99236	0.99108	0.99192	0.99253	0.99068	1	
X_{10}	0.99667	0.99277	0.99067	0.99354	0.99172	0.99326	0.99313	0.99174	0.99217	1

Table 5.9 The value of correlation for $p = 5, n = 30$

$p = 5$	X_1	X_2	X_3	X_4	X_5
X_1	1				
X_2	0.99592	1			
X_3	0.99634	0.99052	1		
X_4	0.99612	0.99141	0.99341	1	
X_5	0.99702	0.99352	0.99002	0.99366	1

Table 5.10 The value of correlation for $p = 2, n = 15$

$p = 2$	X_1	X_2
X_1	1	
X_2	0.99745	1

Table 5.11 The value of correlation for $p = 2, n = 30$

$p = 2$	X_1	X_2
X_1	1	
X_2	0.99411	1

Table 5.12 The value of correlation for $p = 2, n = 50$

$p = 2$	X_1	X_2
X_1	1	
X_2	0.99701	1

These are the simulation results for the MSE values for the three groups shown in tables 5.2, 5.3, 5.4.

Table 5.13 Estimated MSE for group one of simulated data .

n	p	OLS	HKa	KS	HK	KF	Least MSE
15	2	16.986	6.3381	6.2758	6.3278	5.763	KF
	10	468.65	244.49	244.49	290.53	232.75	KF
30	2	7.4152	2.8663	2.8063	2.8607	2.6027	KF
	5	33.936	15.557	15.56	15.844	16.701	HKa
	10	95.501	50.803	50.809	53.717	56.031	HKa
	15	202.92	117.1525	117.1538	133.2	124.03	HKa
50	2	4.2688	1.6641	1.6155	1.6609	1.5105	KF
	10	46.496	24.901	24.914	25.96	27.618	HKa
	20	133.69	80.005	80.01	89.129	86.498	HKa
	30	304.6	183.011	183.028	215.87	197.13	HKa
80	2	2.6138	0.97105	0.93951	0.96955	0.88261	KF
	10	26.097	14.232	14.248	14.798	16.033	HKa
	20	64.895	39.413	39.422	42.124	42.984	HKa
	40	199.77	124.83	124.83	140.68	136.14	HKa
	60	631.14	409.381	409.381	486.53	419.02	HKa

Table 5.14 Estimated MSE for group two of simulated data

n	c	p	OLS	HKa	KS	HK	KF	Higher correlation
100	1	6	0.17108	0.084923	0.12992	0.083226	0.10201	0.71272
		12	0.38877	0.2166	0.26115	0.21579	0.27373	0.74028
		20	0.72207	0.42804	0.46427	0.41937	0.54714	0.74471
		50	3.0193	2.004	2.0146	2.0911	2.4056	0.7914
		70	8.3383	5.4145	5.4172	6.2204	6.4362	0.80601
100	0.5	6	0.49183	0.2472	0.28801	0.25125	0.28152	0.89703
		12	1.1491	0.64544	0.67227	0.65693	0.75272	0.91057
		20	2.1568	1.3315	1.3471	1.3559	1.5142	0.92589
		50	9.0569	6.0566	6.0612	6.7555	6.6275	0.93661
		70	29.432	19.047	19.048	22.463	20.767	0.93716

Table 5.15 Estimated MSE for group three of simulated data

Number of observation	Correlation Between X_1 and X_2	OLS	HKa	KS	HK	KF
15	0.95	2.063	0.7927	0.77804	0.77624	0.71493
	0.9	0.49203	0.20409	0.23834	0.18742	0.17964
	0.8	0.40028	0.17699	0.21265	0.1609	0.15586
	0.7	0.31725	0.13881	0.18424	0.12487	0.12263
30	0.95	0.6661	0.2665	0.2918	0.2066	0.2408
	0.9	0.36352	0.14457	0.1872	0.13735	0.1291
	0.8	0.22678	0.095918	0.13951	0.088443	0.084853
	0.7	0.14291	0.06181	0.10466	0.055694	0.055365
50	0.95	0.52658	0.21188	0.24313	0.20884	0.19193
	0.9	0.21006	0.087371	0.13054	0.083856	0.078623
	0.8	0.12667	0.056689	0.093066	0.052756	0.050665
	0.7	0.083119	0.036533	0.068304	0.032283	0.032143
80	0.95	0.2658	0.10415	0.14896	0.10202	0.094022
	0.9	0.1278	0.054566	0.091562	0.052062	0.048805
	0.8	0.078712	0.033553	0.063701	0.030799	0.029751
	0.7	0.05141	0.022254	0.045328	0.019814	0.019812
100	0.95	0.20773	0.083295	0.12621	0.081379	0.075124
	0.9	0.10266	0.043203	0.077243	0.040924	0.038397
	0.8	0.052756	0.023722	0.045886	0.021714	0.021128
	0.7	0.052617	0.022578	0.045734	0.020669	0.020116

Results in Table 5.13 , when the correlation is too high i.e., when $r = 0.998$ indicating that KF estimator perform better than the other estimators when the number of observations is small i.e when $n = 15$, and for each set of $p = 2$. But HKa perform better for all $n \neq 15$ and $p \neq 2$ of group one. Also we note that as n, p increases HKa and KS perform the same. Moreover, it is observed that for given n and p , the MSEs for all estimators increase as the number of explanatory variables increases.

Results in Table 5.14 , indicating that HK perform better when the correlation is between 0.7 and .81 and for small p , i.e when $p = 6, 12, 20$. But for the same range of correlation we note that HKa perform better for large p , i.e when $p = 50, 70$. When the correlation is between 0.89 and 0.94, we note that HKa perform better than the other estimators for all number of regressors.

Result in Table 5.15, indicating that HKa, KS, HK and KF perform extremely better than the OLS, and KF perform better than the other estimators. Moreover, it is observed that for the given n and p , the MSEs for all estimators decrease as the correlation between regressors decreases.

5.2 Real data

5.2.1 Data base

In order to illustrate the use of ridge regression analysis and assess the potentials of the multiplicative competitive interaction model in the study of shopping behavior. We consider a data set from Leinhardt and Wassermann (1979) which was used in Fox (1997) and is available in the SPSS package.

An automotive industry group keeps track of the sales for a variety of personal motor vehicles. In an effort to be able to identify over and underperforming models, a relationship between vehicle sales and vehicle characteristics need to be established. Data concerning different makes and models of cars is contained in car_sales.sav, see Appendix B for more information. The aim of this application is to use linear regression to identify models that are not selling well.

Nine predictor variables selected for the study are listed in Table 5.16. The response variable is the Sales in thousands (for linearity purpose the Log(Sales) will be considered).

Table 5.16 The selected variables of the vehicle characteristics.

Variable	Variable Name	Description
Y	Ln(sales)	Sales in thousands
X_1	Price	Price in thousands
X_2	engine_s	Engine size
X_3	Horsepow	Horsepower
X_4	Wheelbase	Wheelbase
X_5	Width	Width
X_6	Length	Length
X_7	curb_wgt	Curb weight
X_8	fuel_cap	Fuel capacity
X_9	Mpg	Fuel efficiency

5.2.2 Data analysis

To start the analysis we shall assume that the standard assumptions of the linear regression model hold. A histogram with normal probability plot and P-P plot of the residuals were considered in figures 5.3 and 5.4. The shape of the histogram should approximately follow the shape of the normal curve. This histogram is acceptably close to the normal curve. The P-P plotted residuals should follow the 45-degree line. Neither the histogram nor the P-P plot indicates that the normality assumption is violated.

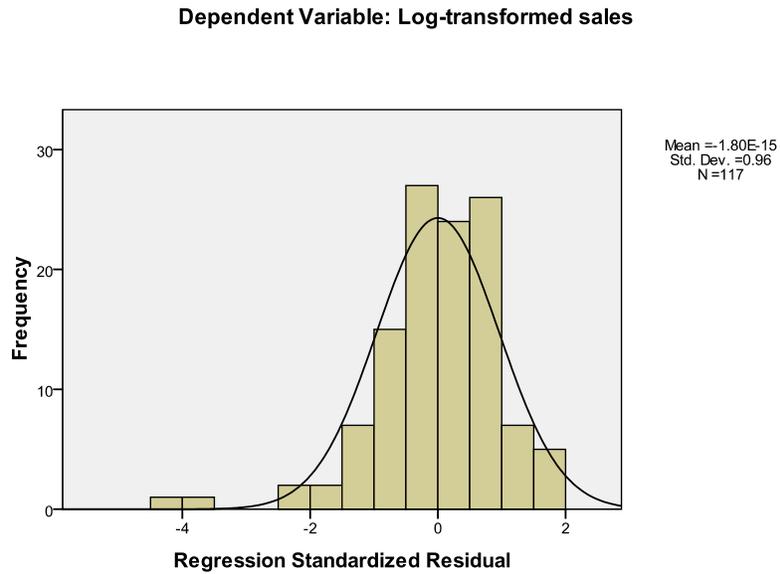


Figure 5.3: Histogram with normal probability plot of the residuals

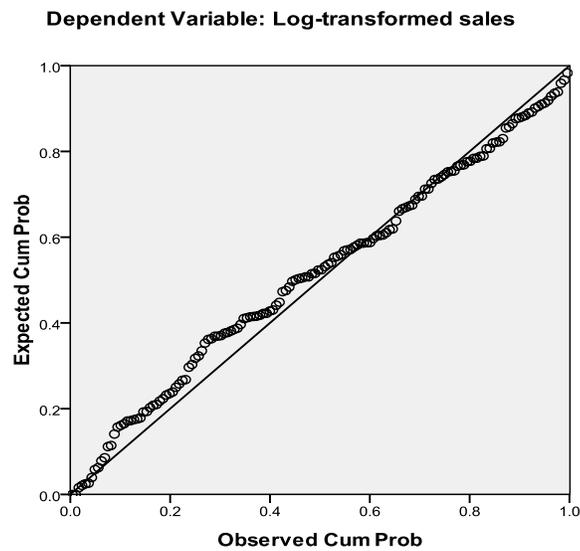


Figure 5.4: Normal P-P Plot of Regression Standardized Residual

As can be expected from the nature of the variables, some of them are highly correlated with each other, results are shown in Table 5.17.

Table 5.17. Correlation Coefficients between deferent variables.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
X_1	1								
X_2	.649**	1							
X_3	.853**	.862**	1						
X_4	.067	.410**	.226*	1					
X_5	.301**	.672**	.507**	.676**	1				
X_6	.183	.537**	.401**	.854**	.743**	1			
X_7	.511**	.743**	.599**	.676**	.736**	.684**	1		
X_8	.406**	.617**	.480**	.659**	.672**	.563**	.848**	1	
X_9	-.480**	-.725**	-.596**	-.471**	-.600**	-.466**	-.819**	-.809**	1

A regression model were fit to the data set. The results were presented in the following tables. The ANOVA table 5.18 reports a significant F statistic (Sig = 0.000), indicating that using the model is better than guessing the mean. A whole, the regression does a good job of modeling sales. Nearly half the variation in sales is explained by the model ($R^2 = 471$) table 5.19.

Table 5.18: Checking the Model Fit (ANOVA)

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	83.285	9	9.254	7.964	.000 ^a
	Residual	124.333	107	1.162		
	Total	207.618	116			

Table 5.19: Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.684 ^a	.467	.431	1.07796

The initial OLS results from fitting a linear model to the data are given in Table 5.20. Although the model fit looks positive. There are several non-significant coefficients, indicating that these variables do not contribute much to the model.

Table 5.20: Model Coefficients

Model parameter	Unstandardized Coefficients		Standardized Coefficients	T	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	-1.301	3.125		-.416	.678		
Price in thousands	-.046	.017	-.489	-2.793	.006	.182	5.487
Engine size	.323	.256	.255	1.264	.209	.138	7.268
Horsepower	-.003	.006	-.124	-.497	.620	.091	11.030
Wheelbase	.092	.030	.553	3.108	.002	.177	5.657
Width	-.027	.052	-.071	-.516	.607	.300	3.337
Length	-.017	.017	-.175	-.968	.335	.171	5.847
Curb weight	.317	.460	.141	.689	.493	.133	7.536
Fuel capacity	-.062	.060	-.176	-1.027	.307	.190	5.258
Fuel efficiency	.029	.048	.095	.599	.551	.223	4.481

The next part of this analysis is to check for multicollinearity. Results in table 5.21 shows that there might be a problem with multicollinearity. For most predictors, the values of the partial and part correlations drop sharply from the zero-order correlation. This means, for example, that much of the variance in sales that is explained by price is also explained by other variables. In collinearity statistics columns, the tolerance is the percentage of the variance in a given predictor that cannot be explained by the other predictors. Thus, the small tolerances show that 70%-90% of the variance in a given predictor can be explained by the other predictors. When the tolerances are close to 0, there is high multicollinearity and the standard error of the regression coefficients will be inflated. A variance inflation

factor greater than 2 is usually considered problematic, and the smallest VIF in table 5.21 is 3.337.

The collinearity diagnostics confirm that there are serious problems with multicollinearity. Several eigenvalues are close to 0, indicating that the predictors are highly intercorrelated and that small changes in the data values may lead to large changes in the estimates of the coefficients. The condition indices are computed as the square roots of the ratios of the largest eigenvalue to each successive eigenvalue. Values greater than 15 indicate a possible problem with collinearity; greater than 30, a serious problem. Six of these indices are larger than 30, suggesting a very serious problem with collinearity.

Table 5.21: Collinearity Diagnostics

	Correlations			Collinearity Statistics		Collinearity Diagnostics	
	Zero-order	Partial	Part	Tolerance	VIF	Eigenvalue	Condition Index
Price in thousands	-.552	-.290	-.217	.187	5.337	.259	6.193
Engine size	-.135	.156	.113	.162	6.159	.050	14.051
Horsepower	-.389	-.043	-.031	.112	8.896	.019	22.589
Wheelbase	.292	.149	.108	.200	4.997	.008	35.942
Width	.037	-.057	-.041	.313	3.193	.005	44.275
Length	.215	.087	.062	.178	5.605	.003	58.480
Curb weight	-.041	.038	.027	.131	7.644	.002	76.175
Fuel capacity	-.016	-.101	-.073	.189	5.303	.001	130.747
Fuel efficiency	.121	.168	.122	.217	4.604	.000	148.267

5.2.3 performance of Ridge Regression to real data

Now, ridge regression will be implemented to fix the collinearity problems. Figure 5.5 illustrates the steps used for finding the best Model for the real data. This Figure shows the steps in ridge regression algorithm that used in this study.

<p>STEP 1 : Center and scale the data</p> $\frac{1}{\sqrt{n-1}} \left(\frac{X-\bar{X}}{\sigma_X} \right) = \tilde{X} \quad , \quad \frac{1}{\sqrt{n-1}} \left(\frac{Y-\bar{Y}}{\sigma_Y} \right) = \tilde{Y}$
<p>STEP 2 : Compute the correlation matrix for centered and scaled data</p> $\tilde{X}^t \tilde{X} = r_{XX}, \quad \tilde{X}^t \tilde{Y} = r_{XY}$
<p>STEP 3 : Compute the ridge parameter k for the values</p> $\hat{k}_{HKa}, \hat{k}_{KS}, \hat{k}_{HK}, \hat{k}_{KF}$
<p>STEP 4 : Compute the ridge regression estimators for the values of k</p> $\tilde{b} = (r_{XX} + kI)^{-1} r_{XY}$
<p>STEP 5 : Compute the coefficients of the natural variables</p> $b_i = \left(\frac{\sigma_Y}{\sigma_X} \right) \tilde{b}_i, \quad i = 1, 2, \dots, p$
<p>STEP 6 : Compute The constant term b_0</p> $b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2 - \dots - b_p \bar{X}_p$
<p>STEP 7 : Choose the model with least mean square</p>

FIGURE 5.5 : Steps in Ridge Regression algorithm

The estimated MSE for the considered ridge method and the OLS is summarized in table 5.22 and the ridge coefficients is listed in table 5.23

5.22 estimated MSEs for real data

OLS	HKa	KS	HK	FK
0.31001	0.22499	0.2393	0.18389	0.26346

Table 5.22 indicating that HKa, KS, HK and FK perform better than the OLS and HK perform extremely better than HKa, KS and FK. Thus our preferred Model that represent the real data is

$$y = b_{HK}X$$

Or

$$\begin{aligned} \text{sales} = & \text{Exp} [-1.50425 - 0.039(\text{price}) + 0.211037(\text{Engine size}) - \\ & 0.00356 (\text{Horsepower}) + 0.066371(\text{Wheelbase}) - 0.01497 (\text{Width}) - \\ & 0.00282 (\text{Length}) + 0.161929 (\text{Curb weight}) - 0.0346 (\text{Fuel capacity}) + \\ & 0.019482 (\text{Fuel efficiency})] \end{aligned}$$

5.23 estimated ridge coefficient for real data

<i>HKa</i>	<i>KS</i>	<i>HK</i>	<i>FK</i>
-1.34806	-1.33246	-1.50425	-1.31321
-0.04364	-0.04422	-0.039	-0.04502
0.290408	0.298967	0.211037	0.310012
-0.00323	-0.00316	-0.00356	-0.00305
0.082317	0.084451	0.066371	0.087431
-0.02332	-0.02418	-0.01497	-0.02529
-0.01131	-0.01253	-0.00282	-0.01425
0.252012	0.265456	0.161929	0.284838
-0.05136	-0.05371	-0.0346	-0.05704
0.025217	0.025976	0.019482	0.027027

5.3 Summary and Conclusions

In this thesis, we studied a comprehensive linear regression models, focusing on the use of ridge regression models performed in a population-based highly correlated data . Analyzes involving such data are quite common in medical, trading, industrial, and various sciences research. The primary goal of such studies may be to simultaneously study the effect of one variable or variables on other variable, but secondary objectives, such as understanding the within variables patterns of correlation, or the relationship between the marker's profiles and the occurrence of the event of interest.

In this research we have studied the properties of a well known approach for choosing the ridge parameter k , when multicollinearity among the columns of the design matrix exists. The investigation has been done using simulated data sets generated from Normal distribution using MatLab v10 software package, also a real data set were considered. In addition to different multicollinearity levels, the number of observation and the number of regressors have been varied. For each combination, we have used 1000 replications. The evaluation of ridge regression approaches has been done by comparing the MSEs among different approaches.

The simulation studies and the analysis of real data set demonstrate that when the correlation is too high i.e., when $r = 0.998$, KF estimator perform better than the other estimators when the number of observations is small i.e when $n = 15$, and for each set with number of observation (15,30 , 50, 80) of $p = 2$. But HKa perform better for all $n \neq 15$ and $p \neq 2$ of group one. Also we note that as n, p increases HKa and KS perform

the same. Moreover, it is observed that for given n and p , the MSEs for all estimators increase as the number of explanatory variables increases.

For group 2 HK perform better when the correlation is between 0.7 and .81 and for small p , i.e when $p = 6, 12, 20$. But for the same range of correlation, we note that HKa perform better for large p , i.e when $p = 50, 70$. When the correlation is between 0.89 and 0.94, we note that HKa perform better than the other estimators for all number of regressors.

For group 3 HKa, KS, HK and KF perform extremely better than the OLS, and KF perform better than the other estimators. Moreover, it is observed that for the given n and p , the MSEs for all estimators decrease as the correlation between regressors decreases.

REFERENCES

- Affi, A.A. and Clark, Y., (1984).** Computer-aided multivariate analysis. Lifetime Learning Publications, Belmont, California.
- A.K.Md. Ehsanes SALEH and B.M. Golam KIBRIA(1993).** Department of Mathematics & Statistics Carleton University, Ottawa, CANADA K1S 5B6.
- Alan T.K. Wan (2002).** Generalized ridge regression estimators under collinearity and balanced loss, Department of Management Sciences, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong.
- Alkhamisi, M. A., Khalaf, G., Shukur, G. (2006).** Some modifications for choosing ridge parameter. Communications in Statistics A 35:1–16.
- Arthur E. Hoerl; Robert W. Kennard Technometrics (Feb., 1970).** Ridge Regression: Biased Estimation for Nonorthogonal Problems Vol. 12, No. 1pp. 55-67.
- Chawla. J.S. (1988).** A note on general ridge estimator, Comm. Statmt.-Theory; Methods 17 (3), 739-744.
- Chawla. J.S. (1989).** The existence theorem in general ridge regression, Statist. Prohah. Lett. 7, 135-137.
- Darlington, R.B. (1978)** “Reduced-Variance Regression,” Psychological Bulletin, , 85, 1238-1255.

Dempster, A.P., Schatzoff, M., & Wermuth, N. (1977) “A Simulation Study of Alternatives to Ordinary Least Squares,” *Journal of the American Statistical Association*, , 72, 77-91.

Douglas Bates(2010). Penalized least squares versus generalized least squares representations of linear mixed models Department of Statistics University of Wisconsin.

Eric Doviak(July 30, 2009). Summary of Ridge Regression.

Farebrother, (1975) “Principal component estimators and minimum mean squares criteria in regression analysis”, *Review of Econometrics and statistics*, 54, 332-336.

Farebrother, R.W. (1978), Partitioned ridge regression. *Technometrics* 20, 121-122.

Galarneau, 1975,*Journal of the American Statistical association* .

GHADBAN KHALAFI AND GHAZI SHUKUR (2005). Choosing Ridge Parameter for Regression Problems, Department of Mathematics, King Khalid University, Saudi Arabia Departments of Economics & Statistics, Jönköping University and Växjö University, Sweden.

G. R. Pasha and Muhammad Akbar Ali Shah (2004). Application of ridge regression to multicollinear data , Bahauddin Zakariya ,University, Multan, Pakistan. Vol.15, No.1, , pp. 97-106 *Journal of Research (Science)*

Hocking, R.R., 1976. The analysis and selection of variables in linear regression. *Biometrics* 32, 1–49.

Hocking, R.R. (1996). Methods and Applications of Linear Models : Regression and the Analysis of Variance. USA : John Wiley & Sons.

Hocking, R.R., Speed, F.M., Lynn, M.J., 1976. A class of biased estimators in linear regression. Technometrics 18 (4), 425–437.

Hoerl, A. E., Kennard, R. W. (1970a). Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 12:55–67.

Hoerl, A. E., Kennard, R. W. (1970b). Ridge regression: application to non-orthogonal problems. Technometrics. 12:69–82.

Hoerl, A. E., Kennard, R. W., Baldwin, K. F. (1975). Ridge regression: some simulations. Commun. Statist. A(4):105–123.

Khalaf, G., Shukur, G., 2005. Choosing ridge parameter for regression problems. Communications in Statistics-Theory and Methods 34,1177–1182.

Kibria, B.M., 2003. Performance of some new ridge regression estimators. Communications in Statistics- Simulation and Computation 32 (2), 419–435.

Lawless and Wang, P. (1976) “A simulation study of Ridge and other Regression Estimators”, Communications in Statistics, Part A-Theory and Method 5, 307-323.

Leinhardt, S. and Wasserman, S. S. (1979) Exploratory data analysis: An introduction to selected methods. In Schuessler, K. (Ed.) Sociological Methodology 1979 Jossey-Bass.

M. A. ALKHAMISI AND G. SHUKUR(2007). A Monte Carlo Study of Recent Ridge Parameters Department of Economics and Statistics, Centre for Labour Market.

Marquardt , D.W. and Snee, R.D. (1970) “Generalized in inverses, Ridge regression, Biased linear estimation”, Tecchnometrics, 12, 591-612.

Marquardt, D.W. and Snee, R.D. (1970) “Ridge regression in practice”. The American Stat., 29, 3-19.

Mallows, H. (1973) “Modern Factor Analysis”, University of Chicago Press.

Miller, A.J., (1990). Subset Selection in Regression. Chapman & Hall, New York.

Myers, R.H., (1986). Classical and Modern Regression With Applications. 2nd Ed.

USA : PWS-KENT Publishing Company.

Neter, J., Kutner, M.H., Nachtseim, C.J. and Wasserman, W., (1996). Applied Linear

Statistical Models. 4th Ed. USA : Irwin Bokk Team.

Neter, J., Wasserman, W. and Kutner, M.H., (1990). Applied Linear Regression

Models. 3rd Ed. USA : IRWIN Book Team.

Neter, J., Wasserman, W. and Kutner, M.H., (1985). Applied Linear Statistical

Models. 2nd Ed. USA : IRWIN.

Paul M. C. de Boer (2005). Ridge regression revisited, Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands.

RANJIT KUMAR PAUL M. Sc. (Agricultural Statistics), Roll No. 4405 I.A.S.R.I,
Library Avenue, New Delhi-110012

Sedlacek, W.E. & Brooks, G.C., Jr. (1976) Racism in American Education: A Model For
Change. Chicago: Nelson-Hall

Vinod. H.D. and A. Ullah (1981). Recent Advances in Regress Methods (Dekker, New
York and Basel).

Wesolowsky, G.O., (1976). Multiple Regression and Analysis of Variance. USA : John
Wiley & Sons.

WALTER A. SHEWHART and SAMUEL S. WILKS (2006). Regression Analysis by
Example, Library of Congress.

Xin Yan , Xiao Gang Su (2009). Linear Regression Analysis Theory and Computing
University of Missouri–Kansas City, USA University of Central Florida.

Yazid M. Al-Hassan(2010). Performance of a new ridge regression estimator Department
of Mathematics, Royal Institute of Technology (KTH), Stockholm, Sweden.

APPENDIX A

This is the data generating function for simulation the three groups (chapter 5)

```
m=1000; ms=zeros(m,5);

for i=1:m,

    n= ; p= ;
    z=zeros(n,p);x=zeros(n,p);y=zeros(n,1);mse_all=zeros(n,1); Ms_ols=0;
    MS_hk=0;

    z(:,1)=randn(n,1);
    for j=2:p,
        z(:,j)=z(:,1)+randn(n,1)*c;
    end
    z;
    for j=1:p,
        x(:,j)=(z(:,j)-mean(z(:,j)))/(((n-1)^(1/2))*std(z(:,j)));
    end
    for j=1:p,
        y1=(sum(z(:,j))')'+randn(n,1);
    end
    y=(y1-mean(y1))/(((n-1)^(1/2))*std(y1));

    rx=x'*x;
    ry=x'*y;
    b=inv(rx)*ry;
    q=(y-x*b)'*(y-x*b)/(n-p);
    ei=eig(rx);
    sumei=0;
    for j=1:p'
        sumei=sumei+(1/ei(j));
    end
    Ms_OLS=q*sumei;
    k1=q/(max(b)^2);
    e=eig(rx);

    sume=0;
    for s=1:p'
        sume=sume+(e(s)/(e(s)+k1)^2);
    end

    MS_HKa=q*sume+k1^2*b'*inv(rx+k1*eye(p,p))*inv(rx+k1*eye(p,p))*b;
    b1=inv(rx+k1*eye(p,p))*ry;

    k2=(q*max(ei))/((max(ei)*(max(b))^2)+(q*(n-p)));
    w=eig(rx);

    sumw=0;
    for d=1:p'
        sumw=sumw+(w(d)/(w(d)+k2)^2);
    end
    b1=inv(rx+k2*eye(p,p))*ry;
```

```

MS_KS=q*sumw+(k2^2)*b'*inv(rx+k2*eye(p,p))*inv(rx+k2*eye(p,p))*b;

K3=(p*q/(b'*b));
e=eig(rx);

sume=0;
for s=1:p'
    sume=sume+(e(s)/(e(s)+k3)^2);
end

MS_HK=q*sume+(k3^2)*b'*inv(rx+k3*eye(p,p))*inv(rx+k3*eye(p,p))*b;
b1=inv(rx+k3*eye(p,p))*ry;
k4=(q/(b'*b));
e=eig(rx);

sume=0;
for s=1:p'
    sume=sume+(e(s)/(e(s)+k4)^2);
end

MS_FK=q*sume+(k4^2)*b'*inv(rx+k4*eye(p,p))*inv(rx+k4*eye(p,p))*b;
b1=inv(rx+k4*eye(p,p))*ry;

ms(i, 1)=Ms_ols; ms(i, 2)=MS_HKa;    ms(i, 3)=MS_KS;
ms(i, 4)=MS_HK; ms(i, 5)=MS_KF;
end

for j=1:5'
means = mean(ms(:,j))

end
MM=( [mean(ms(:,1)) mean(ms(:,2)) mean(ms(:,3)) mean(ms(:,4))
mean(ms(:,5)) ] )

```

APPENDIX B

Table of real data

Ln(sales)	Price	engine_s	horsepow	wheelbas	width	Length	curb_wgt	fuel_cap	mpg
2.83	21.5	1.8	140	101.2	67.3	172.4	2.639	13.2	28
3.67	28.4	3.2	225	108.1	70.3	192.9	3.517	17.2	25
2.15	42	3.5	210	114.6	71.4	196.6	3.85	18	22
3.02	23.99	1.8	150	102.6	68.2	178	2.998	16.4	27
2.93	33.95	2.8	200	108.7	76.1	192	3.561	18.5	22
0.32	62	4.2	310	113	74	198.2	3.902	23.7	21
2.22	33.4	2.8	193	107.3	68.5	176	3.197	16.6	24
2.86	38.9	2.8	193	111.4	70.9	188	3.472	18.5	25
4.52	21.975	3.1	175	109	72.7	194.6	3.368	17.5	25
3.67	25.3	3.8	240	109	72.7	196.2	3.543	17.5	23
3.33	31.965	3.8	205	113.8	74.7	206.8	3.778	18.5	24
4.42	27.885	3.8	205	112.2	73.5	200	3.591	17.5	25
4.15	39.895	4.6	275	115.3	74.5	207.2	3.978	18.5	22
1.88	39.665	4.6	275	108	75.5	200.6	3.843	19	22
2.41	31.01	3	200	107.4	70.3	194.8	3.77	18	22
4.98	13.26	2.2	115	104.1	67.9	180.9	2.676	14.3	27
4.91	16.535	3.1	170	107	69.4	190.4	3.051	15	25
3.2	18.89	3.1	175	107.5	72.5	200.9	3.33	16.6	25
3.75	19.39	3.4	180	110.5	72.7	197.9	3.34	17	27
3.27	24.34	3.8	200	101.1	74.1	193.2	3.5	16.8	25
2.89	45.705	5.7	345	104.5	73.6	179.7	3.21	19.1	22
3.48	13.96	1.8	120	97.1	66.7	174.3	2.398	13.2	33
3.08	9.235	1	55	93.1	62.6	149.4	1.895	10.3	45
2.06	19.84	2.5	163	103.7	69.7	190.9	2.967	15.9	24
3.49	24.495	2.5	168	106	69.2	193	3.332	16	24
3.44	22.245	2.7	200	113	74.4	209.1	3.452	17	26
3.48	16.48	2	132	108	71	186	2.911	16	27
2.6	28.34	3.5	253	113	74.4	207.7	3.564	17	23
4.33	12.64	2	132	105	74.4	174.4	2.567	12.5	29
1.55	19.045	2.5	163	103.7	69.1	190.2	2.879	15.9	24
4.27	20.23	2.5	168	108	71	186	3.058	16	24
-0.09	69.725	8	450	96.2	75.7	176.7	3.375	19	16
5.43	19.46	5.2	230	138.7	79.3	224.2	4.47	26	17
2.82	21.315	3.9	175	109.6	78.8	192.6	4.245	32	15
3.44	18.575	3.9	175	127.2	78.8	208.5	4.298	32	16
4.71	16.98	2.5	120	131	71.5	215	3.557	22	19
5.2	19.565	2.4	150	113.3	76.8	186.3	3.533	20	24
4.25	12.07	2	110	98.4	67	174.7	2.468	12.7	30

4.73	21.56	3.8	190	101.3	73.1	183.2	3.203	15.7	24
3.56	17.035	2.5	170	106.5	69.1	184.6	2.769	15	25
5.5	17.885	3	155	108.5	73	197.6	3.368	16	24
4.15	22.195	4.6	200	114.7	78.2	212	3.908	19	21
5.62	31.93	4	210	111.6	70.2	190.7	3.876	21	19
5.05	21.41	3	150	120.7	76.6	200.9	3.761	26	21
4.83	36.135	4.6	240	119	78.7	204.6	4.808	26	16
5.4	12.05	2.5	119	117.5	69.4	200.7	3.086	20	23
6.29	26.935	4.6	220	138.5	79.1	224.5	4.241	25.1	18
5.3	12.885	1.6	106	103.2	67.1	175.1	2.339	11.9	32
5.44	15.35	2.3	135	106.9	70.3	188.8	2.932	17.1	27
4.29	20.55	2	146	103.2	68.9	177.6	3.219	15.3	24
2.55	26.6	3.2	205	106.4	70.4	178.2	3.857	21.1	19
4.33	26	3.5	210	118.1	75.6	201.2	4.288	20	23
3.72	9.699	1.5	92	96.1	65.7	166.7	2.24	11.9	31
4.2	11.799	2	140	100.4	66.9	174	2.626	14.5	27
3.38	14.999	2.4	148	106.3	71.6	185.4	3.072	17.2	25
3.17	29.465	3	227	108.3	70.2	193.7	3.342	18.5	25
4.02	14.46	2.5	120	93.4	66.7	152	3.045	19	17
4.39	21.62	4	190	101.4	69.4	167.5	3.194	20	20
5.06	26.895	4	195	105.9	72.3	181.5	3.88	20.5	19
3.18	31.505	3	210	105.1	70.5	190.2	3.373	18.5	23
2.54	37.805	3	225	110.2	70.9	189.2	3.638	19.8	23
1.85	54.005	4	290	112.2	72	196.7	3.89	22.5	22
2.62	39.08	4.6	275	109	73.6	208.5	3.868	20	22
3.89	43.33	4.6	215	117.7	78.2	215.3	4.121	19	21
3.27	13.987	1.8	113	98.4	66.5	173.6	2.25	13.2	30
3.75	19.047	2.4	154	100.8	68.9	175.4	2.91	15.9	24
4.02	17.357	2.4	145	103.7	68.5	187.8	2.945	16.3	25
1.74	24.997	3.5	210	107.1	70.3	194.1	3.443	19	22
-2.21	25.45	3	161	97.2	72.4	180.3	3.131	19.8	21
2.43	31.807	3.5	200	107.3	69.9	186.6	4.52	24.3	18
3.67	22.527	3	173	107.3	66.7	178.3	3.51	19.5	20
2.66	16.24	2	125	106.5	69.1	184.8	2.769	15	28
3.28	16.54	2	125	106.4	69.6	185	2.892	16	30
4.22	19.035	3	153	108.5	73	199.7	3.379	16	24
4.4	22.605	4.6	200	114.7	78.2	212	3.958	19	21
3.32	27.56	4	210	111.6	70.2	190.1	3.876	21	18
3.01	22.51	3.3	170	112.2	74.9	194.7	3.944	20	21
2.91	31.75	2.3	185	105.9	67.7	177.4	3.25	16.4	26
3.32	49.9	3.2	221	111.5	70.8	189.4	3.823	21.1	25

2.82	69.7	4.3	275	121.5	73.1	203.1	4.133	23.2	21
1.2	82.6	5	302	99	71.3	177.1	4.125	21.1	20
3.75	13.499	1.8	126	99.8	67.3	177.5	2.593	13.2	30
4.48	20.39	2.4	155	103.1	69.1	183.5	3.012	15.9	25
4.38	26.249	3	222	108.3	70.3	190.5	3.294	18.5	25
3.31	26.399	3.3	170	112.2	74.9	194.8	3.991	20	21
3.75	29.299	3.3	170	106.3	71.7	182.6	3.947	21	19
0.11	18.145	3.1	150	107	69.4	192	3.102	15.2	25
2.69	36.229	4	250	113.8	74.4	205.4	3.967	18.5	22
3	31.598	4.3	190	107	67.8	181.2	4.068	17.5	19
3.19	25.345	3.4	185	120	72.2	201.4	3.948	25	22
3.49	12.64	2	132	105	74.4	174.4	2.559	12.5	29
1.66	16.08	2	132	108	71	186.3	2.942	16	27
3.18	18.85	2.4	150	113.3	76.8	186.3	3.528	20	24
3.94	21.61	2.4	150	104.1	68.4	181.9	2.906	15	27
4.88	19.72	3.4	175	107	70.4	186.3	3.091	15.2	25
2.99	25.31	3.8	200	101.1	74.5	193.4	3.492	16.8	25
4.53	21.665	3.8	195	110.5	72.7	196.5	3.396	18	25
3.58	23.755	3.8	205	112.2	72.6	202.5	3.59	17.5	24
2.2	41.43	2.7	217	95.2	70.1	171	2.778	17	22
0.25	71.02	3.4	300	92.6	69.5	174.5	3.032	17	21
0.62	74.97	3.4	300	92.6	69.5	174.5	3.075	17	23
4.39	10.685	1.9	100	102.4	66.4	176.9	2.332	12.1	33
3.2	12.535	1.9	100	102.4	66.4	180	2.367	12.1	33
1.65	14.29	1.9	124	102.4	66.4	176.9	2.452	12.1	31
4.96	13.108	1.8	120	97	66.7	174	2.42	13.2	33
5.51	17.518	2.2	133	105.2	70.1	188.5	2.998	18.5	27
4.16	25.545	3	210	107.1	71.7	191.9	3.417	18.5	26
3.5	16.875	1.8	140	102.4	68.3	170.5	2.425	14.5	31
4.43	11.528	2.4	142	103.3	66.5	178.7	2.58	15.1	23
3.22	16.888	2	127	94.9	66.7	163.8	2.668	15.3	27
4.23	22.288	2.7	150	105.3	66.5	183.3	3.44	18.5	23
2.29	51.728	4.7	230	112.2	76.4	192.5	5.115	25.4	15
2.28	14.9	2	115	98.9	68.3	163.3	2.767	14.5	26
4.43	16.7	2	115	98.9	68.3	172.3	2.853	14.5	26
3.93	21.2	1.8	150	106.4	68.5	184.1	3.043	16.4	27
2.26	19.99	2	115	97.4	66.7	160.4	3.079	13.7	26
1.72	17.5	2	115	98.9	68.3	163.3	2.762	14.6	26

APPENDIX C

This is the Matlab code performed on real data

```
z=A;
y1=B;
p=9;
n=117
for j=1:9,
    x(:,j)=(z(:,j)-mean(z(:,j)))/(((116)^(1/2))*std(z(:,j)));
end
x;
y=(y1-mean(y1))/(((116)^(1/2))*std(y1));
rx=x'*x;
ry=x'*y;
b=inv(rx)*ry;
q=(y-x*b)'*(y-x*b)/(n-p);
ei=eig(rx);
sumei=0;
for j=1:9'
    sumei=sumei+(1/ei(j));
end
Ms_ols=q*sumei;
k1=q/(max(b)^2);
e=eig(rx);

sume=0;
for s=1:9'
    sume=sume+(e(s)/(e(s)+k1)^2);
end

MS_HKa=q*sume+k1^2*b'*inv(rx+k1*eye(9,9))*inv(rx+k1*eye(9,9))*b;
b1=inv(rx+k1*eye(9,9))*ry;

k2=(q*max(ei))/((max(ei)*(max(b))^2)+(q*(108)));
w=eig(rx);

sumw=0;
for d=1:p'
    sumw=sumw+(w(d)/(w(d)+k2)^2);
end
b2=inv(rx+k2*eye(9,9))*ry;
MS_KS=q*sumw+(k2^2)*b'*inv(rx+k2*eye(9,9))*inv(rx+k2*eye(9,9))*b;
```

```

K3=(p*q/(b'*b));
e=eig(rx);

sume=0;
for s=1:p'
    sume=sume+(e(s)/(e(s)+k3)^2);
end

MS_HK=q*sume+(k3^2)*b'*inv(rx+k3*eye(p,p))*inv(rx+k3*eye(p,p))*b;
b3=inv(rx+k3*eye(p,p))*ry;
k4=(q/(b'*b));
e=eig(rx);

sume=0;
for s=1:p'
    sume=sume+(e(s)/(e(s)+k4)^2);
end

MS_hk2=q*sume+(k4^2)*b'*inv(rx+k4*eye(p,p))*inv(rx+k4*eye(p,p))*b;
b4=inv(rx+k4*eye(p,p))*ry;

M=[Ms_OLS MS_HKa MS_KS MS_HK MS_FK]

where
A=matrix of regressors .

B=vector of dependent variable.

M=vector of MSEs

```